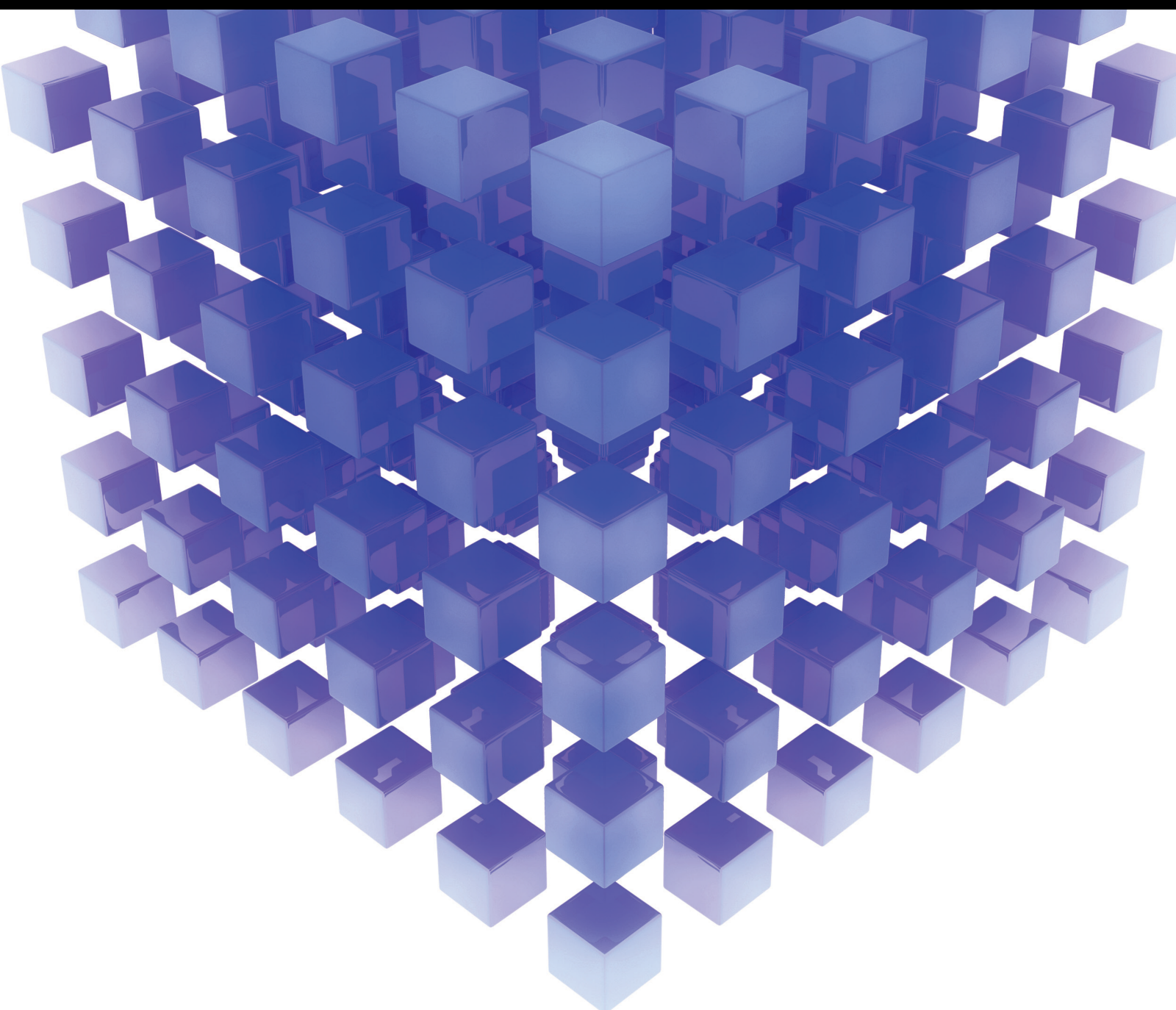


Convergence of Statistical Signal Processing and Machine Learning

Lead Guest Editor: Jie Chen

Guest Editors: Shohel Sayeed and Mohamed Abdellatif





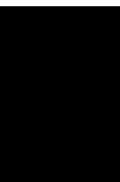
Convergence of Statistical Signal Processing and Machine Learning

Mathematical Problems in Engineering

Convergence of Statistical Signal Processing and Machine Learning

Lead Guest Editor: Jie Chen


Guest Editors: Shohel Sayeed and Mohamed
Abdellatif



Copyright © 2022 Hindawi Limited. All rights reserved.

This is a special issue published in “Mathematical Problems in Engineering.” All articles are open access articles distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Chief Editor

Guangming Xie , China

Academic Editors

Kumaravel A , India
Waqas Abbasi, Pakistan
Mohamed Abd El Aziz , Egypt
Mahmoud Abdel-Aty , Egypt
Mohammed S. Abdo, Yemen
Mohammad Yaghoub Abdollahzadeh
Jamalabadi , Republic of Korea
Rahib Abiyev , Turkey
Leonardo Acho , Spain
Daniela Addessi , Italy
Arooj Adeel , Pakistan
Waleed Adel , Egypt
Ramesh Agarwal , USA
Francesco Aggogeri , Italy
Ricardo Aguilar-Lopez , Mexico
Afaq Ahmad , Pakistan
Naveed Ahmed , Pakistan
Elias Aifantis , USA
Akif Akgul , Turkey
Tareq Al-shami , Yemen
Guido Ala, Italy
Andrea Alaimo , Italy
Reza Alam, USA
Osamah Albahri , Malaysia
Nicholas Alexander , United Kingdom
Salvatore Alfonzetti, Italy
Ghous Ali , Pakistan
Nouman Ali , Pakistan
Mohammad D. Aliyu , Canada
Juan A. Almendral , Spain
A.K. Alomari, Jordan
José Domingo Álvarez , Spain
Cláudio Alves , Portugal
Juan P. Amezcua-Sanchez, Mexico
Mukherjee Amitava, India
Lionel Amodeo, France
Sebastian Anita, Romania
Costanza Arico , Italy
Sabri Arik, Turkey
Fausto Arpino , Italy
Rashad Asharabi , Saudi Arabia
Farhad Aslani , Australia
Mohsen Asle Zaem , USA

Andrea Avanzini , Italy
Richard I. Avery , USA
Viktor Avrutin , Germany
Mohammed A. Awadallah , Malaysia
Francesco Aymerich , Italy
Sajad Azizi , Belgium
Michele Bacciocchi , Italy
Seungik Baek , USA
Khaled Bahlali, France
M.V.A Raju Bahubalendruni, India
Pedro Balaguer , Spain
P. Balasubramaniam, India
Stefan Balint , Romania
Ines Tejado Balsera , Spain
Alfonso Banos , Spain
Jerzy Baranowski , Poland
Tudor Barbu , Romania
Andrzej Bartoszewicz , Poland
Sergio Baselga , Spain
S. Caglar Baslamisli , Turkey
David Bassir , France
Chiara Bedon , Italy
Azeddine Beghdadi, France
Andriette Bekker , South Africa
Francisco Beltran-Carbajal , Mexico
Abdellatif Ben Makhlof , Saudi Arabia
Denis Benasciutti , Italy
Ivano Benedetti , Italy
Rosa M. Benito , Spain
Elena Benvenuti , Italy
Giovanni Berselli, Italy
Michele Betti , Italy
Pietro Bia , Italy
Carlo Bianca , France
Simone Bianco , Italy
Vincenzo Bianco, Italy
Vittorio Bianco, Italy
David Bigaud , France
Sardar Muhammad Bilal , Pakistan
Antonio Bilotta , Italy
Sylvio R. Bistafa, Brazil
Chiara Boccaletti , Italy
Rodolfo Bontempo , Italy
Alberto Borboni , Italy
Marco Bortolini, Italy

Paolo Boscariol, Italy
Daniela Boso , Italy
Guillermo Botella-Juan, Spain
Abdesselem Boulkroune , Algeria
Boulaïd Boulkroune, Belgium
Fabio Bovenga , Italy
Francesco Braghin , Italy
Ricardo Branco, Portugal
Julien Bruchon , France
Matteo Bruggi , Italy
Michele Brun , Italy
Maria Elena Bruni, Italy
Maria Angela Butturi , Italy
Bartłomiej Błachowski , Poland
Dhanamjayulu C , India
Raquel Caballero-Águila , Spain
Filippo Cacace , Italy
Salvatore Caddemi , Italy
Zuowei Cai , China
Roberto Caldelli , Italy
Francesco Cannizzaro , Italy
Maosen Cao , China
Ana Carpio, Spain
Rodrigo Carvajal , Chile
Caterina Casavola, Italy
Sara Casciati, Italy
Federica Caselli , Italy
Carmen Castillo , Spain
Inmaculada T. Castro , Spain
Miguel Castro , Portugal
Giuseppe Catalanotti , United Kingdom
Alberto Cavallo , Italy
Gabriele Cazzulani , Italy
Fatih Vehbi Celebi, Turkey
Miguel Cerrolaza , Venezuela
Gregory Chagnon , France
Ching-Ter Chang , Taiwan
Kuei-Lun Chang , Taiwan
Qing Chang , USA
Xiaoheng Chang , China
Prasenjit Chatterjee , Lithuania
Kacem Chehdi, France
Peter N. Cheimets, USA
Chih-Chiang Chen , Taiwan
He Chen , China
























Kebing Chen , China
Mengxin Chen , China
Shyi-Ming Chen , Taiwan
Xizhong Chen , Ireland
Xue-Bo Chen , China
Zhiwen Chen , China
Qiang Cheng, USA
Zeyang Cheng, China
Luca Chiapponi , Italy
Francisco Chicano , Spain
Tirivanhu Chinyoka , South Africa
Adrian Chmielewski , Poland
Seongim Choi , USA
Gautam Choubey , India
Hung-Yuan Chung , Taiwan
Yusheng Ci, China
Simone Cinquemani , Italy
Roberto G. Citarella , Italy
Joaquim Ciurana , Spain
John D. Clayton , USA
Piero Colajanni , Italy
Giuseppina Colicchio, Italy
Vassilios Constantoudis , Greece
Enrico Conte, Italy
Alessandro Contento , USA
Mario Cools , Belgium
Gino Cortellessa, Italy
Carlo Cosentino , Italy
Paolo Crippa , Italy
Erik Cuevas , Mexico
Guozeng Cui , China
Mehmet Cunkas , Turkey
Giuseppe D'Aniello , Italy
Peter Dabnichki, Australia
Weizhong Dai , USA
Zhifeng Dai , China
Purushothaman Damodaran , USA
Sergey Dashkovskiy, Germany
Adiel T. De Almeida-Filho , Brazil
Fabio De Angelis , Italy
Samuele De Bartolo , Italy
Stefano De Miranda , Italy
Filippo De Monte , Italy

José António Fonseca De Oliveira
Correia , Portugal
Jose Renato De Sousa , Brazil
Michael Defoort, France
Alessandro Della Corte, Italy
Laurent Dewasme , Belgium
Sanku Dey , India
Gianpaolo Di Bona , Italy
Roberta Di Pace , Italy
Francesca Di Puccio , Italy
Ramón I. Diego , Spain
Yannis Dimakopoulos , Greece
Hasan Dinçer , Turkey
José M. Domínguez , Spain
Georgios Dounias, Greece
Bo Du , China
Emil Dumic, Croatia
Madalina Dumitriu , United Kingdom
Premraj Durairaj , India
Saeed Eftekhari Azam, USA
Said El Kafhali , Morocco
Antonio Elipse , Spain
R. Emre Erkmen, Canada
John Escobar , Colombia
Leandro F. F. Miguel , Brazil
FRANCESCO FOTI , Italy
Andrea L. Facci , Italy
Shahla Faisal , Pakistan
Giovanni Falsone , Italy
Hua Fan, China
Jianguang Fang, Australia
Nicholas Fantuzzi , Italy
Muhammad Shahid Farid , Pakistan
Hamed Faruqi, Iran
Yann Favennec, France
Fiorenzo A. Fazzolari , United Kingdom
Giuseppe Fedele , Italy
Roberto Fedele , Italy
Baowei Feng , China
Mohammad Ferdows , Bangladesh
Arturo J. Fernández , Spain
Jesus M. Fernandez Oro, Spain
Francesco Ferrise, Italy
Eric Feulvarch , France
Thierry Floquet, France

Eric Florentin , France
Gerardo Flores, Mexico
Antonio Forcina , Italy
Alessandro Formisano, Italy
Francesco Franco , Italy
Elisa Francomano , Italy
Juan Frausto-Solis, Mexico
Shujun Fu , China
Juan C. G. Prada , Spain
HECTOR GOMEZ , Chile
Matteo Gaeta , Italy
Mauro Gaggero , Italy
Zoran Gajic , USA
Jaime Gallardo-Alvarado , Mexico
Mosè Gallo , Italy
Akemi Gálvez , Spain
Maria L. Gandarias , Spain
Hao Gao , Hong Kong
Xingbao Gao , China
Yan Gao , China
Zhiwei Gao , United Kingdom
Giovanni Garcea , Italy
José García , Chile
Harish Garg , India
Alessandro Gasparetto , Italy
Stylianos Georgantzinou, Greece
Fotios Georgiades , India
Parviz Ghadimi , Iran
Ştefan Cristian Gherghina , Romania
Georgios I. Giannopoulos , Greece
Agathoklis Giaralis , United Kingdom
Anna M. Gil-Lafuente , Spain
Ivan Giorgio , Italy
Gaetano Giunta , Luxembourg
Jefferson L.M.A. Gomes , United Kingdom
Emilio Gómez-Déniz , Spain
Antonio M. Gonçalves de Lima , Brazil
Qunxi Gong , China
Chris Goodrich, USA
Rama S. R. Gorla, USA
Veena Goswami , India
Xunjie Gou , Spain
Jakub Grabski , Poland

Antoine Grall , France
George A. Gravvanis , Greece
Fabrizio Greco , Italy
David Greiner , Spain
Jason Gu , Canada
Federico Guarracino , Italy
Michele Guida , Italy
Muhammet Gul , Turkey
Dong-Sheng Guo , China
Hu Guo , China
Zhaoxia Guo, China
Yusuf Gurefe, Turkey
Salim HEDDAM , Algeria
ABID HUSSANAN, China
Quang Phuc Ha, Australia
Li Haitao , China
Petr Hájek , Czech Republic
Mohamed Hamdy , Egypt
Muhammad Hamid , United Kingdom
Renke Han , United Kingdom
Weimin Han , USA
Xingsi Han, China
Zhen-Lai Han , China
Thomas Hanne , Switzerland
Xinan Hao , China
Mohammad A. Hariri-Ardebili , USA
Khalid Hattaf , Morocco
Defeng He , China
Xiao-Qiao He, China
Yanchao He, China
Yu-Ling He , China
Ramdane Hedjar , Saudi Arabia
Jude Hemanth , India
Reza Hemmati, Iran
Nicolae Herisanu , Romania
Alfredo G. Hernández-Díaz , Spain
M.I. Herreros , Spain
Eckhard Hitzer , Japan
Paul Honeine , France
Jaromir Horacek , Czech Republic
Lei Hou , China
Yingkun Hou , China
Yu-Chen Hu , Taiwan
Yunfeng Hu, China
Can Huang , China
Gordon Huang , Canada
Linsheng Huo , China
Sajid Hussain, Canada
Asier Ibeas , Spain
Orest V. Iftime , The Netherlands
Przemyslaw Ignaciuk , Poland
Giacomo Innocenti , Italy
Emilio Insfran Pelozo , Spain
Azeem Irshad, Pakistan
Alessio Ishizaka, France
Benjamin Ivorra , Spain
Breno Jacob , Brazil
Reema Jain , India
Tushar Jain , India
Amin Jajarmi , Iran
Chiranjibe Jana , India
Łukasz Jankowski , Poland
Samuel N. Jator , USA
Juan Carlos Jáuregui-Correa , Mexico
Kandasamy Jayakrishna, India
Reza Jazar, Australia
Khalide Jbilou, France
Isabel S. Jesus , Portugal
Chao Ji , China
Qing-Chao Jiang , China
Peng-fei Jiao , China
Ricardo Fabricio Escobar Jiménez , Mexico
Emilio Jiménez Macías , Spain
Maolin Jin, Republic of Korea
Zhuo Jin, Australia
Ramash Kumar K , India
BHABEN KALITA , USA
MOHAMMAD REZA KHEDMATI , Iran
Viacheslav Kalashnikov , Mexico
Mathiyalagan Kalidass , India
Tamas Kalmar-Nagy , Hungary
Rajesh Kaluri , India
Jyottheswara Reddy Kalvakurthi, India
Zhao Kang , China
Ramani Kannan , Malaysia
Tomasz Kapitaniak , Poland
Julius Kaplunov, United Kingdom
Konstantinos Karamanos, Belgium
Michal Kawulok, Poland

Irfan Kaymaz , Turkey
Vahid Kayvanfar , Qatar
Krzysztof Kecik , Poland
Mohamed Khader , Egypt
Chaudry M. Khalique , South Africa
Mukhtaj Khan , Pakistan
Shahid Khan , Pakistan
Nam-Il Kim, Republic of Korea
Philipp V. Kiryukhantsev-Korneev ,
Russia
P.V.V Kishore , India
Jan Koci , Czech Republic
Ioannis Kostavelis , Greece
Sotiris B. Kotsiantis , Greece
Frederic Kratz , France
Vamsi Krishna , India
Edyta Kucharska, Poland
Krzysztof S. Kulpa , Poland
Kamal Kumar, India
Prof. Ashwani Kumar , India
Michal Kunicki , Poland
Cedrick A. K. Kwuimy , USA
Kyandoghere Kyamakya, Austria
Ivan Kyrchei , Ukraine
Márcio J. Lacerda , Brazil
Eduardo Lalla , The Netherlands
Giovanni Lancioni , Italy
Jaroslaw Latalski , Poland
Hervé Laurent , France
Agostino Lauria , Italy
Aimé Lay-Ekuakille , Italy
Nicolas J. Leconte , France
Kun-Chou Lee , Taiwan
Dimitri Lefebvre , France
Eric Lefevre , France
Marek Lefik, Poland
Yaguo Lei , China
Kauko Leiviskä , Finland
Ervin Lenzi , Brazil
ChenFeng Li , China
Jian Li , USA
Jun Li , China
Yueyang Li , China
Zhao Li , China































Zhen Li , China
En-Qiang Lin, USA
Jian Lin , China
Qibin Lin, China
Yao-Jin Lin, China
Zhiyun Lin , China
Bin Liu , China
Bo Liu , China
Heng Liu , China
Jianxu Liu , Thailand
Lei Liu , China
Sixin Liu , China
Wanquan Liu , China
Yu Liu , China
Yuanchang Liu , United Kingdom
Bonifacio Llamazares , Spain
Alessandro Lo Schiavo , Italy
Jean Jacques Loiseau , France
Francesco Lolli , Italy
Paolo Lonetti , Italy
António M. Lopes , Portugal
Sebastian López, Spain
Luis M. López-Ochoa , Spain
Vassilios C. Loukopoulos, Greece
Gabriele Maria Lozito , Italy
Zhiguo Luo , China
Gabriel Luque , Spain
Valentin Lychagin, Norway
YUE MEI, China
Junwei Ma , China
Xuanlong Ma , China
Antonio Madeo , Italy
Alessandro Magnani , Belgium
Toqeer Mahmood , Pakistan
Fazal M. Mahomed , South Africa
Arunava Majumder , India
Sarfranz Nawaz Malik, Pakistan
Paolo Manfredi , Italy
Adnan Maqsood , Pakistan
Muazzam Maqsood, Pakistan
Giuseppe Carlo Marano , Italy
Damijan Markovic, France
Filipe J. Marques , Portugal
Luca Martinelli , Italy
Denizar Cruz Martins, Brazil

Francisco J. Martos , Spain
Elio Masciari , Italy
Paolo Massioni , France
Alessandro Mauro , Italy
Jonathan Mayo-Maldonado , Mexico
Pier Luigi Mazzeo , Italy
Laura Mazzola, Italy
Driss Mehdi , France
Zahid Mehmood , Pakistan
Roderick Melnik , Canada
Xiangyu Meng , USA
Jose Merodio , Spain
Alessio Merola , Italy
Mahmoud Mesbah , Iran
Luciano Mescia , Italy
Laurent Mevel , France
Constantine Michailides , Cyprus
Mariusz Michta , Poland
Prankul Middha, Norway
Aki Mikkola , Finland
Giovanni Minafò , Italy
Edmondo Minisci , United Kingdom
Hiroyuki Mino , Japan
Dimitrios Mitsotakis , New Zealand
Ardashir Mohammadzadeh , Iran
Francisco J. Montáns , Spain
Francesco Montefusco , Italy
Gisele Mophou , France
Rafael Morales , Spain
Marco Morandini , Italy
Javier Moreno-Valenzuela , Mexico
Simone Morganti , Italy
Caroline Mota , Brazil
Aziz Moukrim , France
Shen Mouquan , China
Dimitris Mourtzis , Greece
Emiliano Mucchi , Italy
Taseer Muhammad, Saudi Arabia
Ghulam Muhiuddin, Saudi Arabia
Amitava Mukherjee , India
Josefa Mula , Spain
Jose J. Muñoz , Spain
Giuseppe Muscolino, Italy
Marco Mussetta , Italy

Hariharan Muthusamy, India
Alessandro Naddeo , Italy
Raj Nandkeolyar, India
Keivan Navaie , United Kingdom
Soumya Nayak, India
Adrian Neagu , USA
Erivelton Geraldo Nepomuceno , Brazil
AMA Neves, Portugal
Ha Quang Thinh Ngo , Vietnam
Nhon Nguyen-Thanh, Singapore
Papakostas Nikolaos , Ireland
Jelena Nikolic , Serbia
Tatsushi Nishi, Japan
Shanzhou Niu , China
Ben T. Nohara , Japan
Mohammed Nouari , France
Mustapha Nourelfath, Canada
Kazem Nouri , Iran
Ciro Núñez-Gutiérrez , Mexico
Włodzimierz Ogryczak, Poland
Roger Ohayon, France
Krzysztof Okarma , Poland
Mitsuhiro Okayasu, Japan
Murat Olgun , Turkey
Diego Oliva, Mexico
Alberto Olivares , Spain
Enrique Onieva , Spain
Calogero Orlando , Italy
Susana Ortega-Cisneros , Mexico
Sergio Ortobelli, Italy
Naohisa Otsuka , Japan
Sid Ahmed Ould Ahmed Mahmoud , Saudi Arabia
Taoreed Owolabi , Nigeria
EUGENIA PETROPOULOU , Greece
Arturo Pagano, Italy
Madhumangal Pal, India
Pasquale Palumbo , Italy
Dragan Pamučar, Serbia
Weifeng Pan , China
Chandan Pandey, India
Rui Pang, United Kingdom
Jürgen Pannek , Germany
Elena Panteley, France
Achille Paolone, Italy

George A. Papakostas , Greece
Xosé M. Pardo , Spain
You-Jin Park, Taiwan
Manuel Pastor, Spain
Pubudu N. Pathirana , Australia
Surajit Kumar Paul , India
Luis Payá , Spain
Igor Pažanin , Croatia
Libor Pekař , Czech Republic
Francesco Pellicano , Italy
Marcello Pellicciari , Italy
Jian Peng , China
Mingshu Peng, China
Xiang Peng , China
Xindong Peng, China
Yuxing Peng, China
Marzio Pennisi , Italy
Maria Patrizia Pera , Italy
Matjaz Perc , Slovenia
A. M. Bastos Pereira , Portugal
Wesley Peres, Brazil
F. Javier Pérez-Pinal , Mexico
Michele Perrella, Italy
Francesco Pesavento , Italy
Francesco Petrini , Italy
Hoang Vu Phan, Republic of Korea
Lukasz Pieczonka , Poland
Dario Piga , Switzerland
Marco Pizzarelli , Italy
Javier Plaza , Spain
Goutam Pohit , India
Dragan Poljak , Croatia
Jorge Pomares , Spain
Hiram Ponce , Mexico
Sébastien Poncet , Canada
Volodymyr Ponomaryov , Mexico
Jean-Christophe Ponsart , France
Mauro Pontani , Italy
Sivakumar Poruran, India
Francesc Pozo , Spain
Aditya Rio Prabowo , Indonesia
Anchasa Pramuanjaroenkij , Thailand
Leonardo Primavera , Italy
B Rajanarayan Prusty, India

Krzysztof Puszynski , Poland
Chuan Qin , China
Dongdong Qin, China
Jianlong Qiu , China
Giuseppe Quaranta , Italy
DR. RITU RAJ , India
Vitomir Racic , Italy
Carlo Rainieri , Italy
Kumbakonam Ramamani Rajagopal, USA
Ali Ramazani , USA
Angel Manuel Ramos , Spain
Higinio Ramos , Spain
Muhammad Afzal Rana , Pakistan
Muhammad Rashid, Saudi Arabia
Manoj Rastogi, India
Alessandro Rasulo , Italy
S.S. Ravindran , USA
Abdolrahman Razani , Iran
Alessandro Reali , Italy
Jose A. Reinoso , Spain
Oscar Reinoso , Spain
Haijun Ren , China
Carlo Renno , Italy
Fabrizio Renno , Italy
Shahram Rezapour , Iran
Ricardo Rianza , Spain
Francesco Riganti-Fulginei , Italy
Gerasimos Rigatos , Greece
Francesco Ripamonti , Italy
Jorge Rivera , Mexico
Eugenio Roanes-Lozano , Spain
Ana Maria A. C. Rocha , Portugal
Luigi Rodino , Italy
Francisco Rodríguez , Spain
Rosana Rodríguez López, Spain
Francisco Rossomando , Argentina
Jose de Jesus Rubio , Mexico
Weiguo Rui , China
Rubén Ruiz , Spain
Ivan D. Rukhlenko , Australia
Dr. Eswaramoorthi S. , India
Weichao SHI , United Kingdom
Chaman Lal Sabharwal , USA
Andrés Sáez , Spain

Bekir Sahin, Turkey
Laxminarayan Sahoo , India
John S. Sakellariou , Greece
Michael Sakellariou , Greece
Salvatore Salamone, USA
Jose Vicente Salcedo , Spain
Alejandro Salcido , Mexico
Alejandro Salcido, Mexico
Nunzio Salerno , Italy
Rohit Salgotra , India
Miguel A. Salido , Spain
Sinan Salih , Iraq
Alessandro Salvini , Italy
Abdus Samad , India
Sovan Samanta, India
Nikolaos Samaras , Greece
Ramon Sancibrian , Spain
Giuseppe Sanfilippo , Italy
Omar-Jacobo Santos, Mexico
J Santos-Reyes , Mexico
José A. Sanz-Herrera , Spain
Musavarah Sarwar, Pakistan
Shahzad Sarwar, Saudi Arabia
Marcelo A. Savi , Brazil
Andrey V. Savkin, Australia
Tadeusz Sawik , Poland
Roberta Sburlati, Italy
Gustavo Scaglia , Argentina
Thomas Schuster , Germany
Hamid M. Sedighi , Iran
Mijanur Rahaman Seikh, India
Tapan Senapati , China
Lotfi Senhadji , France
Junwon Seo, USA
Michele Serpilli, Italy
Silvestar Šesnić , Croatia
Gerardo Severino, Italy
Ruben Sevilla , United Kingdom
Stefano Sfarra , Italy
Dr. Ismail Shah , Pakistan
Leonid Shaikhet , Israel
Vimal Shanmuganathan , India
Prayas Sharma, India
Bo Shen , Germany
Hang Shen, China

Xin Pu Shen, China
Dimitri O. Shepelsky, Ukraine
Jian Shi , China
Amin Shokrollahi, Australia
Suzanne M. Shontz , USA
Babak Shotorban , USA
Zhan Shu , Canada
Angelo Sifaleras , Greece
Nuno Simões , Portugal
Mehakpreet Singh , Ireland
Piyush Pratap Singh , India
Rajiv Singh, India
Seralathan Sivamani , India
S. Sivasankaran , Malaysia
Christos H. Skiadas, Greece
Konstantina Skouri , Greece
Neale R. Smith , Mexico
Bogdan Smolka, Poland
Delfim Soares Jr. , Brazil
Alba Sofi , Italy
Francesco Soldovieri , Italy
Raffaele Solimene , Italy
Yang Song , Norway
Jussi Sopanen , Finland
Marco Spadini , Italy
Paolo Spagnolo , Italy
Ruben Specogna , Italy
Vasilios Spitas , Greece
Ivanka Stamova , USA
Rafał Stanisławski , Poland
Miladin Stefanović , Serbia
Salvatore Strano , Italy
Yakov Strelniker, Israel
Kangkang Sun , China
Qiuqin Sun , China
Shuaishuai Sun, Australia
Yanchao Sun , China
Zong-Yao Sun , China
Kumarasamy Suresh , India
Sergey A. Suslov , Australia
D.L. Suthar, Ethiopia
D.L. Suthar , Ethiopia
Andrzej Swierniak, Poland
Andras Szekrenyes , Hungary
Kumar K. Tamma, USA

Yong (Aaron) Tan, United Kingdom
Marco Antonio Taneco-Hernández , Mexico
Lu Tang , China
Tianyou Tao, China
Hafez Tari , USA
Alessandro Tasora , Italy
Sergio Teggi , Italy
Adriana del Carmen Téllez-Anguiano , Mexico
Ana C. Teodoro , Portugal
Efstathios E. Theotokoglou , Greece
Jing-Feng Tian, China
Alexander Timokha , Norway
Stefania Tomasiello , Italy
Gisella Tomasini , Italy
Isabella Torricollo , Italy
Francesco Tornabene , Italy
Mariano Torrisi , Italy
Thang nguyen Trung, Vietnam
George Tsiatas , Greece
Le Anh Tuan , Vietnam
Nerio Tullini , Italy
Emilio Turco , Italy
Ilhan Tuzcu , USA
Efstratios Tzirtzilakis , Greece
FRANCISCO UREÑA , Spain
Filippo Ubertini , Italy
Mohammad Uddin , Australia
Mohammad Safi Ullah , Bangladesh
Serdar Ulubeyli , Turkey
Mati Ur Rahman , Pakistan
Panayiotis Vafeas , Greece
Giuseppe Vairo , Italy
Jesus Valdez-Resendiz , Mexico
Eusebio Valero, Spain
Stefano Valvano , Italy
Carlos-Renato Vázquez , Mexico
Martin Velasco Villa , Mexico
Franck J. Vernerey, USA
Georgios Veronis , USA
Vincenzo Vespri , Italy
Renato Vidoni , Italy
Venkatesh Vijayaraghavan, Australia

Anna Vila, Spain
Francisco R. Villatoro , Spain
Francesca Vipiana , Italy
Stanislav Vitek , Czech Republic
Jan Vorel , Czech Republic
Michael Vynnycky , Sweden
Mohammad W. Alomari, Jordan
Roman Wan-Wendner , Austria
Bingchang Wang, China
C. H. Wang , Taiwan
Dagang Wang, China
Guoqiang Wang , China
Huaiyu Wang, China
Hui Wang , China
J.G. Wang, China
Ji Wang , China
Kang-Jia Wang , China
Lei Wang , China
Qiang Wang, China
Qingling Wang , China
Weiwei Wang , China
Xinyu Wang , China
Yong Wang , China
Yung-Chung Wang , Taiwan
Zhenbo Wang , USA
Zhibo Wang, China
Waldemar T. Wójcik, Poland
Chi Wu , Australia
Qihong Wu, China
Yuqiang Wu, China
Zhibin Wu , China
Zhizheng Wu , China
Michalis Xenos , Greece
Hao Xiao , China
Xiao Ping Xie , China
Qingzheng Xu , China
Binghan Xue , China
Yi Xue , China
Joseph J. Yame , France
Chuanliang Yan , China
Xinggang Yan , United Kingdom
Hongtai Yang , China
Jixiang Yang , China
Mijia Yang, USA
Ray-Yeng Yang, Taiwan

Zaoli Yang , China
Jun Ye , China
Min Ye , China
Luis J. Yebra , Spain
Peng-Yeng Yin , Taiwan
Muhammad Haroon Yousaf , Pakistan
Yuan Yuan, United Kingdom
Qin Yuming, China
Elena Zaitseva , Slovakia
Arkadiusz Zak , Poland
Mohammad Zakwan , India
Ernesto Zambrano-Serrano , Mexico
Francesco Zammori , Italy
Jessica Zangari , Italy
Rafal Zdunek , Poland
Ibrahim Zeid, USA
Nianyin Zeng , China
Junyong Zhai , China
Hao Zhang , China
Haopeng Zhang , USA
Jian Zhang , China
Kai Zhang, China
Lingfan Zhang , China
Mingjie Zhang , Norway
Qian Zhang , China
Tianwei Zhang , China
Tongqian Zhang , China
Wenyu Zhang , China
Xianming Zhang , Australia
Xuping Zhang , Denmark
Yinyan Zhang, China
Yifan Zhao , United Kingdom
Debao Zhou, USA
Heng Zhou , China
Jian G. Zhou , United Kingdom
Junyong Zhou , China
Xueqian Zhou , United Kingdom
Zhe Zhou , China
Wu-Le Zhu, China
Gaetano Zizzo , Italy
Mingcheng Zuo, China

Contents

Kernel Least Logarithmic Absolute Difference Algorithm

Dongliang Fu, Wei Gao , Wentao Shi, and Qunfei Zhang

Research Article (11 pages), Article ID 9092663, Volume 2022 (2022)

Radar-Based Rainfall Estimation of Landfalling Tropical Storm “PABUK” 2019 over Southern Thailand

Pakdee Chantraket , Sukrit Kirtsaeng , Chakrit Chaotamonsak, Somporn Chantara , Supachai Nakapan, and Thammarat Panityakul 



Research Article (11 pages), Article ID 9968329, Volume 2022 (2022)

Application of HMC-SS Method in Pipeline Reliability Analysis and Residual Life Assessment

Xin Lin  and Guojian Shao 

Research Article (10 pages), Article ID 3756441, Volume 2021 (2021)

Fault Diagnosis of Rolling Bearing Based on Improved VMD and KNN

Quanbo Lu , Xinqi Shen, Xiujun Wang, Mei Li , Jia Li, and Mengzhou Zhang

Research Article (11 pages), Article ID 2530315, Volume 2021 (2021)

Modeling and Design of an Aircraft-Mode Controller for a Fixed-Wing VTOL UAV

Zhao Deng , Liaoni Wu, and Yancheng You



Research Article (17 pages), Article ID 7902134, Volume 2021 (2021)

Adversarial Sample Detection with Gaussian Mixture Conditional Generative Adversarial Networks

Pengfei Zhang  and Xiaoming Ju 





Research Article (18 pages), Article ID 8268249, Volume 2021 (2021)

Cooperative Target Search of UAV Swarm with Communication Distance Constraint

Ning Wang , Zhe Li , Xiaolong Liang, Ying Li, and Feihu Zhao


Research Article (14 pages), Article ID 3794329, Volume 2021 (2021)

A Low-Order Partial Integrated Guidance and Control Scheme for Diving Hypersonic Vehicles to Impact Ground Maneuver Target

Tong An , JianHua Wang , YuLong Pan , and HaiShan Chen 


Research Article (12 pages), Article ID 7407739, Volume 2021 (2021)

Analysis of Assembly Tolerance Based on Assembly Constraint Information Model

Chunxi Li  and Wenjun Hou


Research Article (18 pages), Article ID 7438966, Volume 2021 (2021)

High Efficiency Spam Filtering: A Manifold Learning-Based Approach

Chao Wang , Qun Li, Tian-yu Ren, Xiao-hu Wang, and Guang-xin Guo

Research Article (7 pages), Article ID 2993877, Volume 2021 (2021)

Numerical Investigation of the Trailing Edge Shape on the Added Damping of a Kaplan Turbine Runner

Zhang Ming, P. A. Mbango-Ngoma, Du Xiao-zhen, and Chen Qing-Guang 

Research Article (11 pages), Article ID 9559454, Volume 2021 (2021)

Research on Fault Diagnosis of Launch Vehicle's Power Transformation and Transmission System Based on Big Data

Yichi Zhang , Tao Shu, Xincheng Song, Yan Xu, and Pengxiang Zhang

Research Article (11 pages), Article ID 3108000, Volume 2021 (2021)

Fire Prediction Based on CatBoost Algorithm

Fangrong Zhou, Hao Pan, Zhenyu Gao, Xuyong Huang, Guochao Qian, Yu Zhu, and Feng Xiao 





Research Article (9 pages), Article ID 1929137, Volume 2021 (2021)

Prediction of Financial Time Series Based on LSTM Using Wavelet Transform and Singular Spectrum Analysis

Qi Tang , Ruchen Shi , Tongmei Fan , Yidan Ma , and Jingyan Huang 

Research Article (13 pages), Article ID 9942410, Volume 2021 (2021)

Prefabricated Concrete Component Geometry Deviation Statistical Analysis

Xiaoyong Luo , Hao Long , Shuang Dong , and Jingyi Wu 

Research Article (15 pages), Article ID 9993451, Volume 2021 (2021)

Research Article

Kernel Least Logarithmic Absolute Difference Algorithm

Dongliang Fu,¹ Wei Gao ,² Wentao Shi,³ and Qunfei Zhang³

¹Shanghai Marine Equipment Research Institute, Shanghai 200031, China

²School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang 212013, China

³School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an 710072, China

Correspondence should be addressed to Wei Gao; wei_gao@ujs.edu.cn

Received 2 July 2021; Revised 8 December 2021; Accepted 31 January 2022; Published 8 March 2022

Academic Editor: Akif Akgul

Copyright © 2022 Dongliang Fu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Kernel adaptive filtering (KAF) algorithms derived from the second moment of error criterion perform very well in nonlinear system identification under assumption of the Gaussian observation noise; however, they inevitably suffer from severe performance degradation in the presence of non-Gaussian impulsive noise and interference. To resolve this dilemma, we propose a novel robust kernel least logarithmic absolute difference (KLLAD) algorithm based on logarithmic error cost function in reproducing kernel Hilbert spaces, taking into account of the non-Gaussian impulsive noise. The KLLAD algorithm shows considerable improvement over the existing KAF algorithms without restraining impulsive interference in terms of robustness and convergence speed. Moreover, the convergence condition of KLLAD algorithm with Gaussian kernel and fixed dictionary is presented in the mean sense. The superior performance of KLLAD algorithm is confirmed by the simulation results.

1. Introduction

Kernel adaptive filters as a tremendous breakthrough of the conventional linear adaptive filters have been widely used in many practical nonlinear applications including time series prediction [1], acoustic echo cancellation [2], channel equalization [3], abnormal event detection [4], etc. The scheme of kernel adaptive filtering (KAF) is to map the original input data into high or infinite dimensional feature space via kernel function and then apply the framework of typical linear adaptive filtering to the transformed data in the reproducing kernel Hilbert spaces (RKHS) leading to various KAF algorithms [5–8]. The kernel least-mean-square (KLMS) algorithm, as the benchmark among of KAF algorithms, is developed from the cost function of second-order statistic of the error between the desired signal and instantaneous estimate under the assumption of Gaussian noise for its mathematical simplicity and convenience [9]. Therefore, it can be ensured that the performances of KLMS-type algorithms only for Gaussian disturbance noise severely degrade, when the desired signals are corrupted by the impulsive interferences. In practical applications, e.g., underwater acoustic signal processing [10], wireless

communication environments [11], and radar cluster elimination [12], the impulsive noises with the statistical characteristics consisting of infrequency, short duration, and high amplitude are more rational than the ideal Gaussian noises.

For conventional linear robust adaptive filtering, the signed adaptive filters and their theoretical analyses have been extensively studied in [13–15]. The generalized maximum correntropy criterion (GMCC) algorithm with generalized Gaussian density function was proposed in [16], and its stability and steady-state mean square performance were also investigated. The authors of [17] proposed the least logarithmic absolute difference (LLAD) with the correntropy-induced metric (CIM) constraint in order to exploit system sparsity and suppress the interferences. The minimum kernel risk-sensitive loss (MKRSL) algorithm was developed to achieve fast convergence speed by applying risk-sensitive loss while yielding the robust performance to outliers [18]. When the shape parameter is equal to 2 and the risk-sensitive parameter gradually tends to trivial, both GMCC and MKRSL algorithms reduce to the ordinary maximum correntropy criterion (MCC) algorithm. The constrained least mean logarithmic square (CLMLS) based

on a relative logarithmic cost function and its variants were proposed in [19], and they were used in the application of sparse sensor array synthesis achieving the desired beam pattern with much less sensor elements. In [20], a robust least mean logarithmic square (RLMLS) algorithm and its variable step-size variant were presented to combat impulsive noises, and its theoretical mean square performance was also analyzed with the stationary white Gaussian inputs.

Meanwhile, the topic of nonlinear system identification corrupted by the impulsive noise also attracted large numbers of research interests [21]. It is thus important and necessary to study the robust KAF algorithm in order to resolve the divergence problem of performance caused by non-Gaussian impulsive noise. In [22], the kernel maximum correntropy criterion (KMCC) algorithm was developed by introducing the framework of MCC into RKHS in the presence of impulsive noises. By combining the kernel method and the affine projection sign algorithm, the kernel affine projection sign algorithm (KAPSA) was proposed in [23] to combat the non-Gaussian impulse interference. An improved variable forgetting factor recursive logarithmic least mean p th (IVFF-RLLMP) algorithm was developed in [24] for the Volterra system identification against the impulsive interference modeled by α -stable distribution. As a counterpart of GMCC algorithm, the generalized kernel maximum correntropy (GKMC) and the quantized GKMC (QGKMC) algorithm were developed in [25] for robust nonlinear adaptive filtering. In [26], the quantized minimum kernel risk-sensitive loss (QMKRSL) algorithm was proposed to achieve better and robust performance of nonlinear filtering for outliers. Motivated by the studies in [27, 28] on the Cauchy loss which has been successfully used in various robust learning applications, the multikernel minimum Cauchy kernel loss (MKMCKL) algorithm was reported in [29] showing the improved nonlinear filtering performance over counterpart single algorithm in the presence of extreme outliers. Recently, the kernel affine projection-like (KAPL) algorithm in RKHS was proposed and investigated for nonlinear channel equalization in scenarios of non-Gaussian noises [30]. The kernel least mean p -power (KLMP) algorithm was proposed to alleviate the adverse impact of impulsive noise in [31, 32], independently. More recently, Nyström kernel recursive generalized maximum correntropy (NKRGMCC) with probability density rank-based quantization sampling algorithm was proposed to improve the convergence performance for impulsive noises [33]. The tracking analysis of kernel signed error algorithm (KSEA) with Gaussian kernel for time-variant nonlinear system was analyzed thoroughly under the assumption of impulsive noise [34]. More recently, the logarithmic hyperbolic cosine-based adaptive filter (LHCAF) was proposed in [35] to address the issue of instability of its prototype algorithm, and the transient and steady-state analyses were also provided. Subsequently, the authors of [36] proposed the multiple random Fourier features Cauchy-loss conjugate gradient (MRFCCG) algorithm which has better performance than the classical KAF algorithms in terms of computational complexity and filtering accuracy.

Therefore, the cost functions adopting the frameworks of fractional order statistics of error or the distinct types of error measures are able to provide effective ways to reveal the robust performance against impulsive noises. Inspired by the family of linear adaptive filtering algorithms based on the logarithmic cost proposed in [37], our purpose is to extend this scheme into RKHS to obtain the robustness of KAF algorithm particularly within non-Gaussian impulsive noise environment. In this paper, the kernel least logarithmic absolute difference algorithm based on the logarithmic error cost framework is proposed to achieve the nonlinear system identification in the impulsive interference environments, which are more frequently encountered in practical applications. Simulation results illustrate the proposed KLLAD algorithm can consistently decrease the drastic perturbation of recursive weight coefficients caused by the large amplitude of instantaneous estimation error with low probability.

Notation: We use normal font small letters x for scalar variables, boldface small letters \mathbf{x} for column vectors, and boldface capital letters \mathbf{X} for matrices. The superscript $(\cdot)^T$ represents the transpose of a vector or a matrix. The expectation is denoted by $\mathbb{E}\{\cdot\}$, and matrix trace is denoted by $\text{tr}\{\cdot\}$. The Gaussian distribution with mean μ and variance σ^2 is denoted by $\mathcal{N}(\mu, \sigma^2)$. The notation $\|\cdot\|_2$ is the ℓ_2 -norm of its matrix or vector argument; $\|\mathbf{x}\|_A^2$ is the weighted square value $\mathbf{x}^T \mathbf{A} \mathbf{x}$. Notation $\text{sgn}\{\cdot\}$ is the signum function. The operator $\text{eig}_{\max}\{\mathbf{X}\}$ denotes the maximum eigenvalue of matrix \mathbf{X} . Identity matrix of size $N \times N$ is denoted by \mathbf{I}_N .

2. Preliminaries of KAF Algorithms

Let \mathcal{H} denote a Hilbert space of real-valued function ψ from a compact subspace $\mathcal{U} \subset \mathbb{R}^L$. Function $\kappa: \mathcal{U} \times \mathcal{U} \rightarrow \mathbb{R}$ is a reproducing kernel, and $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ is the induced RKHS with its inner product. The following unknown nonlinear system is considered:

$$y_n = f^*(\mathbf{x}_n) + z_n. \quad (1)$$

The scalar y_n and the vector $\mathbf{x}_n \in \mathbb{R}^L$ denote the desired output and the input signal, respectively. Here, f^* represents the optimum functional to be identified, and z_n stands for the non-Gaussian impulsive noise modeled as the contaminated-Gaussian (CG) noise as follows [13, 38]:

$$z_n = v_n + b_n \epsilon_n, \quad (2)$$

where both v_n and ϵ_n are statistically independent zero-mean white Gaussian noises with the variances σ_v^2 and $\sigma_\epsilon^2 = K\sigma_v^2$, with parameter $1 \ll K$. Moreover, the random sequence b_n is from a Bernoulli random process with the probability of $\Pr(b_n = 1) = p_r$ and $\Pr(b_n = 0) = 1 - p_r$. Notice that v_n and $b_n \epsilon_n$ represent the common Gaussian noise and particular impulsive interference component, respectively. Subsequently, the probability density function (pdf) of the impulsive CG noise z_n is given by

$$p_z = (1 - p_r) \mathcal{N}(0, \sigma_v^2) + p_r \mathcal{N}(0, (K + 1)\sigma_v^2). \quad (3)$$

When $p_r = 0$, the impulsive CG noise z_n deteriorates to a common white Gaussian noise with zero-mean and variance

σ_v^2 , namely, impulsive interference free. Correspondingly, the variance of impulsive noise z_n is given by

$$\sigma_z^2 = \mathbb{E}\{z_n^2\} = (1 - p_r)\sigma_v^2 + p_r(K + 1)\sigma_v^2. \quad (4)$$

Hence, z_n is statistically independent of any other signals. The reason why we adopt the CG model is that it makes the analysis of mean stability of KLLAD algorithm mathematically tractable.

Given a sample set of pairs of input vectors and desired output scalars, i.e., $\{\mathbf{x}_n, y_n\}_{n=1}^N$, we aim at estimating a nonlinear regression function ψ that relates input data \mathbf{x}_n and output data y_n corrupted by impulsive noise z_n . Let \mathcal{H} be a RKHS with kernel κ . By virtue of the representer theorem [39], the function ψ in \mathcal{H} that minimizes the regularized least-squares problem:

$$\min_{\psi \in \mathcal{H}} \sum_{n=1}^N |y_n - \psi(\mathbf{x}_n)|^2 + \varepsilon \|\psi\|_{\mathcal{H}}^2, \quad (5)$$

with ε a nonnegative regularization constant, can be written as the kernel expansion in terms of available training data in form of

$$\psi = \sum_{n=1}^N w_n \kappa(\cdot, \mathbf{x}_n). \quad (6)$$

When $\varepsilon = 0$, it will not affect the derivation of algorithm. It can be seen from (5) and (6) that the functional representation is formulated as the parametric vector form. However, solution (6) is infeasible to be performed in an online manner, because the algorithm can not cope with the linear increase of the size N of the model as the latest input data available. A commonly used strategy is to adopt a finite-order model of the form [5]

$$\bar{\psi} = \sum_{m=1}^{M_n} w_m \kappa(\cdot, \mathbf{x}_{\omega, m}). \quad (7)$$

The set of selected input data $\mathcal{D}_n = \{\mathbf{x}_{\omega, m}\}_{m=1}^{M_n}$ is so-called online dictionary with the length M_n and is generated by the informative criteria from input vectors \mathbf{x}_n in an online manner, e.g., coherence criterion [1], surprise criterion [5], and approximate linear dependency [40]. Note that (6) seems to be the same as (7) in form, whereas the length M_n determined by the online sparsification criterion is analogue to the order of transversal filter. Without loss of generality, we chose the coherence criterion which allows the candidate \mathbf{x}_n to be inserted into the dictionary if its maximum coherence remains below the given threshold δ_κ , namely,

$$\max_{m=1,2,\dots,M_n} |\kappa(\mathbf{x}_n, \mathbf{x}_{\omega, m})| \leq \delta_\kappa, \quad (8)$$

where $0 < \delta_\kappa \leq 1$ determines both the level of sparsity and the coherence of the dictionary. Based on the stochastic gradient of cost functions $J(e_n)$ with respect to \mathbf{w} , using the parametric finite-order model (7), we can then obtain the corresponding recursive update equation of the KAF algorithms:

$$\mathbf{w}_{n+1} = \mathbf{w}_n + \eta \frac{\partial J(e_n)}{\partial \mathbf{w}} \mathbf{\kappa}_{\omega, n}. \quad (9)$$

With the positive step-size η and the weight coefficients vector $\mathbf{w}_n = [w_n(1), w_n(2), \dots, w_n(M_n)]^\top$, the instantaneous estimation error e_n is given by

$$e_n = y_n - \mathbf{w}_n^\top \mathbf{\kappa}_{\omega, n}. \quad (10)$$

Moreover, the kernelized input vector $\mathbf{\kappa}_{\omega, n}$ is defined by

$$\mathbf{\kappa}_{\omega, n} = [\kappa(\mathbf{x}_n, \mathbf{x}_{\omega, 1}), \kappa(\mathbf{x}_n, \mathbf{x}_{\omega, 2}), \dots, \kappa(\mathbf{x}_n, \mathbf{x}_{\omega, M_n})]^\top. \quad (11)$$

Adopting the mean-squared error (MSE) criterion for (9) leads to the scheme of KLMS algorithm described as follows:

$$\mathbf{w}_{n+1} = \mathbf{w}_n + \eta e_n \mathbf{\kappa}_{\omega, n}, \quad (12)$$

where the complete KLMS algorithm with online sparsification criterion is not present for clarity.

It can be easily observed from (12) that the KLMS-type algorithms suffer from severe performance degradation and even divergence caused by the instantaneous estimation error given in (10), which is contaminated by the non-Gaussian impulsive noise defined in (2). It is thus critical for the design of robust KAF algorithm to effectively suppress the adverse effects of impulsive interference noise.

3. KLLAD Algorithm

In this section, we shall derive the KLLAD algorithm based on the logarithmic cost function.

Logarithmic cost as one type of relative cost measures is capable of providing relatively legitimate amplification for the ordinary error fluctuation and significant attenuation for very large error value induced by impulsive interferences. It has been illustrated in [37] that the logarithmic cost function can proportionally adjust the weight coefficients for small and large error values depending on the combination weights varying with time. Therefore, we introduce the differentiable combined logarithmic error cost function as follows:

$$J(e_n) = F(e_n) - \frac{1}{\lambda} \ln(1 + \lambda F(e_n)), \quad (13)$$

where λ is a design parameter, and $F(e_n)$ is the conventional cost function of instantaneous estimation error. Note that the universal adaptive filtering algorithm obtained from (13) is able to update the coefficients by small error and suppress the drastic perturbation of error simultaneously.

Taking the derivative of (13) with respect to weight vector \mathbf{w} , we obtain

$$\nabla_{\mathbf{w}} J(e_n) = \nabla_{\mathbf{w}} F(e_n) \frac{\lambda F(e_n)}{1 + \lambda F(e_n)}. \quad (14)$$

Since the impulsive CG noise exists, the kernel adaptive filters based on quadratic cost function have to face the severe performance degradation. Nevertheless, the signed adaptive filtering algorithms based on the ℓ_1 -norm of

estimation error are robust to the impulsive noise because of its recursive update equation only utilizing the sign of instantaneous estimation error. In order to mitigate the adverse impact of impulsive noise on KAF, let the conventional cost function of error $F(e_n)$ be $\mathbb{E}\{|e_n|\}$ using (10), namely,

$$F(e_n) = \mathbb{E}\{|y_n - \mathbf{w}_n^\top \boldsymbol{\kappa}_{\omega,n}|\}. \quad (15)$$

It should be pointed out that cost function (13) using (15) can perform less attenuation for the small estimation errors and more attenuation for relatively large estimation errors simultaneously. Substituting (15) into (14) leads to

$$\frac{\partial J(e_n)}{\partial \mathbf{w}} = \frac{\partial F(e_n)}{\partial \mathbf{w}} \frac{\lambda F(e_n)}{1 + \lambda F(e_n)}. \quad (16)$$

Applying the steepest-descent method to minimize the logarithmic cost function (13), thus the weights vector of KLLAD algorithm can be solved iteratively by

$$\mathbf{w}_{n+1} = \mathbf{w}_n + \eta \left[-\frac{\partial J(e_n)}{\partial \mathbf{w}} \right]. \quad (17)$$

By (16), the recursive update equation of KLLAD algorithm can be reformulated as

$$\mathbf{w}_{n+1} = \mathbf{w}_n - \eta \frac{\partial F(e_n)}{\partial \mathbf{w}} \frac{\lambda F(e_n)}{1 + \lambda F(e_n)}, \quad (18)$$

where the subgradient in the second term on the right hand side of (18) is calculated as

$$\frac{\partial F(e_n)}{\partial \mathbf{w}} \approx \text{sgn}(e_n). \quad (19)$$

Note that the above approximation notation implies that the subgradient of conventional cost function $F(e_n)$ is replaced by its instantaneous estimate. Substituting (15) and (19) into (18) and then removing the expectations, the stochastic update equation of KLLAD algorithm can be rewritten as

$$\begin{aligned} \mathbf{w}_{n+1} &= \mathbf{w}_n + \eta \frac{\lambda |e_n|}{1 + \lambda |e_n|} \text{sgn}(e_n) \boldsymbol{\kappa}_{\omega,n} \\ &= \mathbf{w}_n + \eta \frac{\lambda e_n}{1 + \lambda |e_n|} \boldsymbol{\kappa}_{\omega,n}, \end{aligned} \quad (20)$$

where $\boldsymbol{\kappa}_{\omega,n}$ denotes the $M_n \times 1$ dimensional kernelized input vector as in (11). For universal nonlinear filtering performance, we only focus on using the Gaussian kernel function $\kappa(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|_2^2 / 2\xi^2)$, with $\xi > 0$ the kernel bandwidth in this paper. Moreover, the Gaussian kernel function has been successfully used in the theoretical analysis of KAF algorithms due to its mathematical convenience and tractability in the derivation [34, 41]. Based on (20), the recursive update equations of KLLAD algorithm based on the finite-order model using the coherence criterion are presented in the following.

At each time instant n , the input \mathbf{x}_n will be decided into the case of rejection or reception according to the coherence criterion (8) for online dictionary \mathcal{D}_n as follows:

(i) Rejection: if $\max_{m=1,2,\dots,M_n} |\kappa(\mathbf{x}_n, \mathbf{x}_{\omega,m})| > \delta_\kappa$

$$\begin{aligned} \mathcal{D}_{n+1} &= \mathcal{D}_n, M_{n+1} = M_n, \\ \mathbf{w}_{n+1} &= \mathbf{w}_n + \eta \frac{\lambda e_n}{1 + \lambda |e_n|} \boldsymbol{\kappa}_{\omega,n}. \end{aligned} \quad (21)$$

Note that the dimensions of kernelized input vector $\boldsymbol{\kappa}_{\omega,n}$ and weight vector \mathbf{w}_n remain still in this case.

(ii) Reception: if $\max_{m=1,2,\dots,M_n} |\kappa(\mathbf{x}_n, \mathbf{x}_{\omega,m})| \leq \delta_\kappa$

$$\begin{aligned} \mathcal{D}_{n+1} &= \mathcal{D}_n \cup \{\mathbf{x}_n\}, M_{n+1} = M_n + 1, \\ \mathbf{w}_{n+1} &= \begin{bmatrix} \mathbf{w}_n \\ 0 \end{bmatrix} + \eta \frac{\lambda e_n}{1 + \lambda |e_n|} \boldsymbol{\kappa}_{\omega,n}. \end{aligned} \quad (22)$$

Note that the dimensions of kernelized input vector $\boldsymbol{\kappa}_{\omega,n}$ and weight vector \mathbf{w}_n are augmented to make the update equation of weight vector valid.

It should be pointed that the threshold δ_κ is selected by grid search over the interval $[0, 1)$ to determine the sparsity of online dictionary \mathcal{D}_n as in [1, 5–7].

The KLLAD algorithm described by (21) and (22) combines the advantages of KLMS and KSEA algorithms leading to the improved convergence performance taking into account Gaussian noise and non-Gaussian impulsive noise existing simultaneously. When the design parameter λ is set to large value, the KLLAD algorithm can achieve the robust convergence performance against the impulsive noise. In addition, the KLLAD algorithm has faster convergence rate than the KSEA algorithm especially for highly correlated input signals. The scheme of KLLAD algorithm is summarized as Algorithm 1.

By (9), the recursive update equation of KAF algorithms can be expressed in a general form of

$$\mathbf{w}_{n+1} = \mathbf{w}_n + \eta f(e_n) e_n \boldsymbol{\kappa}_{\omega,n}, \quad (23)$$

where $f(e_n)$ is defined as the nonlinear error function which performs like the generalized variable step-size against non-Gaussian impulsive noise. According to (23), various nonlinear KAF algorithms can be readily obtained by solving distinct cost functions in the RKHS. Hence, the error functions of some KAF algorithms mentioned previously are listed in Table 1.

It should be pointed out that the error function of MKRSL algorithm reduces to that of KMCC algorithm as the risk-sensitive parameter $\lambda_f \rightarrow 0_+$ or that of KLMS algorithm as the kernel bandwidth $\xi_f \rightarrow \infty$ [18]. This means that the KMCC algorithm is a particular case of MKRSL algorithm in fact. Thus, we are only concerned with the KMCC algorithm with the characteristic of robustness for impulsive noises in the following. Correspondingly, the curves of error functions listed in Table 1 with different parameters are plotted in Figure 1. First of all, it can be observed from Figure 1(a) that the value of error function of KLMS algorithm is constant one without any ability of suppressing the impulsive noise due to its quadratic cost

- (1) **Input:** $\{\mathbf{x}_n, y_n\}, n = 1, 2, \dots$
- (2) **Initialization:** select the step-size $\eta > 0$, the Gaussian kernel bandwidth $\xi > 0$, the threshold $\delta_\kappa > 0$, the parameter $\lambda > 0$, the dictionary $\mathcal{D}_0 = \{\mathbf{x}_1\}$, and weight coefficient $w_0(1) = 0$.
- (3) **for** $n = 1, 2, \dots$, **do**
- (4) **if** $\max_{m=1,2,\dots,M_n} |\kappa(\mathbf{x}_n, \mathbf{x}_{\omega,m})| > \delta_\kappa$ Update \mathbf{w}_{n+1} via (21);
- (5) **elseif** $\max_{m=1,2,\dots,M_n} |\kappa(\mathbf{x}_n, \mathbf{x}_{\omega,m})| \leq \delta_\kappa$ Update \mathbf{w}_{n+1} via (22);
- (6) **end if** Obtain solution $\bar{\psi}(\mathbf{x}_n) = \sum_{m=1}^{M_n} w_n(m) \kappa(\mathbf{x}_n, \mathbf{x}_{\omega,m})$.
- (7) **end for**

ALGORITHM 1: KLLAD algorithm.

TABLE 1: Error functions of several KAF algorithms.

Algorithm	Error function $f(e_n)$
KLMS	$f(e_n) = 1$
KMCC	$f(e_n) = \kappa(y_n, \phi(\mathbf{x}_n)) = \exp(-\ e_n\ _2^2 / 2\xi_f^2)$
KLMP	$f(e_n) = e_n ^{p-2}$
KLLAD	$f(e_n) = \lambda/1 + \lambda e_n $

function for Gaussian noise. We can then see from Figures 1(b)–1(d) that the output values of nonlinear error functions of KMCC, KLMP, and KLLAD algorithms are effectively attenuated even for the very large instantaneous estimation errors caused by impulsive interferences. Specifically, the attenuation rate of error function of KMCC with small kernel bandwidth ξ_f is very fast, whereas its maximum output values corresponding to the relatively small error inputs particularly near zero are always less than 1. Consequently, the KMCC algorithm has to take much larger step-size η than the KLMP and KLLAD algorithms. Although both error functions have similar shape of output values for KLMP algorithm with small p and KLLAD algorithm with large λ , the error function of KLLAD consistently gives much smaller outputs for larger error inputs than that of KLMP.

4. Convergence Condition on Mean Stability of the KLLAD Algorithm

In this section, we shall investigate the sufficient convergence condition of KLLAD algorithm via Gaussian kernel function with the fixed dictionary in the mean sense.

For mathematical tractability of theoretical analysis, the CG modeled impulsive noise z_n given in (1), instead of the standard symmetric α -stable distribution, has been successfully used in the theoretical analysis of robust adaptive filters [13, 14, 34].

One the one hand, the coherence criterion guarantees the length M_n is infinite [1]. On the other hand, it is true that the length M_n of online dictionary \mathcal{D}_n gradually tends to be invariant in the steady-state phase. Therefore, we only need to consider the recursive update equation (21) of KLLAD algorithm with the steady fixed dictionary \mathcal{D} with constant length M in the context of the derivation of convergence condition on mean stability.

We start with introducing the weight error vector defined by

$$\mathbf{v}_n = \mathbf{w}_n - \mathbf{w}^*, \quad (24)$$

where $\mathbf{w}^* = [w_1^*, w_2^*, \dots, w_M^*]^\top$ represents the optimal weight vector of vector-valued form of nonlinear system. It should be emphasized that \mathbf{w}^* corresponds to the functional f^* based on the specific dictionary elements. From (1) and (24), the instantaneous estimation error (10) can be rewritten as

$$e_n = z_n - \mathbf{v}_n^\top \bar{\mathbf{k}}_{\omega,n}. \quad (25)$$

With the kernelized input vector with fixed dictionary \mathcal{D}

$$\bar{\mathbf{k}}_{\omega,n} = [\kappa(\mathbf{x}_n, \mathbf{x}_{\omega,1}), \kappa(\mathbf{x}_n, \mathbf{x}_{\omega,2}), \dots, \kappa(\mathbf{x}_n, \mathbf{x}_{\omega,M})]^\top. \quad (26)$$

Subtracting \mathbf{w}^* from both sides of (21), and using (25), we obtain

$$\mathbf{v}_{n+1} = \mathbf{v}_n + \eta\lambda \frac{z_n - \mathbf{v}_n^\top \bar{\mathbf{k}}_{\omega,n}}{1 + \lambda|e_n|} \bar{\mathbf{k}}_{\omega,n}. \quad (27)$$

Taking the expected values of both sides of (27), it follows that

$$\begin{aligned} \mathbb{E}\{\mathbf{v}_{n+1}\} &= \mathbb{E}\{\mathbf{v}_n\} + \eta\lambda \mathbb{E}\left\{ \frac{z_n \bar{\mathbf{k}}_{\omega,n}}{1 + \lambda|e_n|} \right\} \\ &\quad - \eta\lambda \mathbb{E}\left\{ \frac{\bar{\mathbf{k}}_{\omega,n} \bar{\mathbf{k}}_{\omega,n}^\top}{1 + \lambda|e_n|} \mathbf{v}_n \right\}. \end{aligned} \quad (28)$$

Since the impulsive CG noise z_n with zero-mean is assumed to be statistically independent of any other signals as mentioned above, and $\bar{\mathbf{k}}_{\omega,n} \bar{\mathbf{k}}_{\omega,n}^\top$ is assumed to be statistically independent of weight error vector \mathbf{v}_n , i.e., the modified independence assumption (MIA) widely used in [34, 41, 42], then (28) can be reformulated as

$$\mathbb{E}\{\mathbf{v}_{n+1}\} = \left(\mathbf{I}_M - \eta\lambda \mathbb{E}\left\{ \frac{\bar{\mathbf{k}}_{\omega,n} \bar{\mathbf{k}}_{\omega,n}^\top}{1 + \lambda|e_n|} \right\} \right) \mathbb{E}\{\mathbf{v}_n\}. \quad (29)$$

For theoretical analysis tractability, the second term of (29) can be approximated as

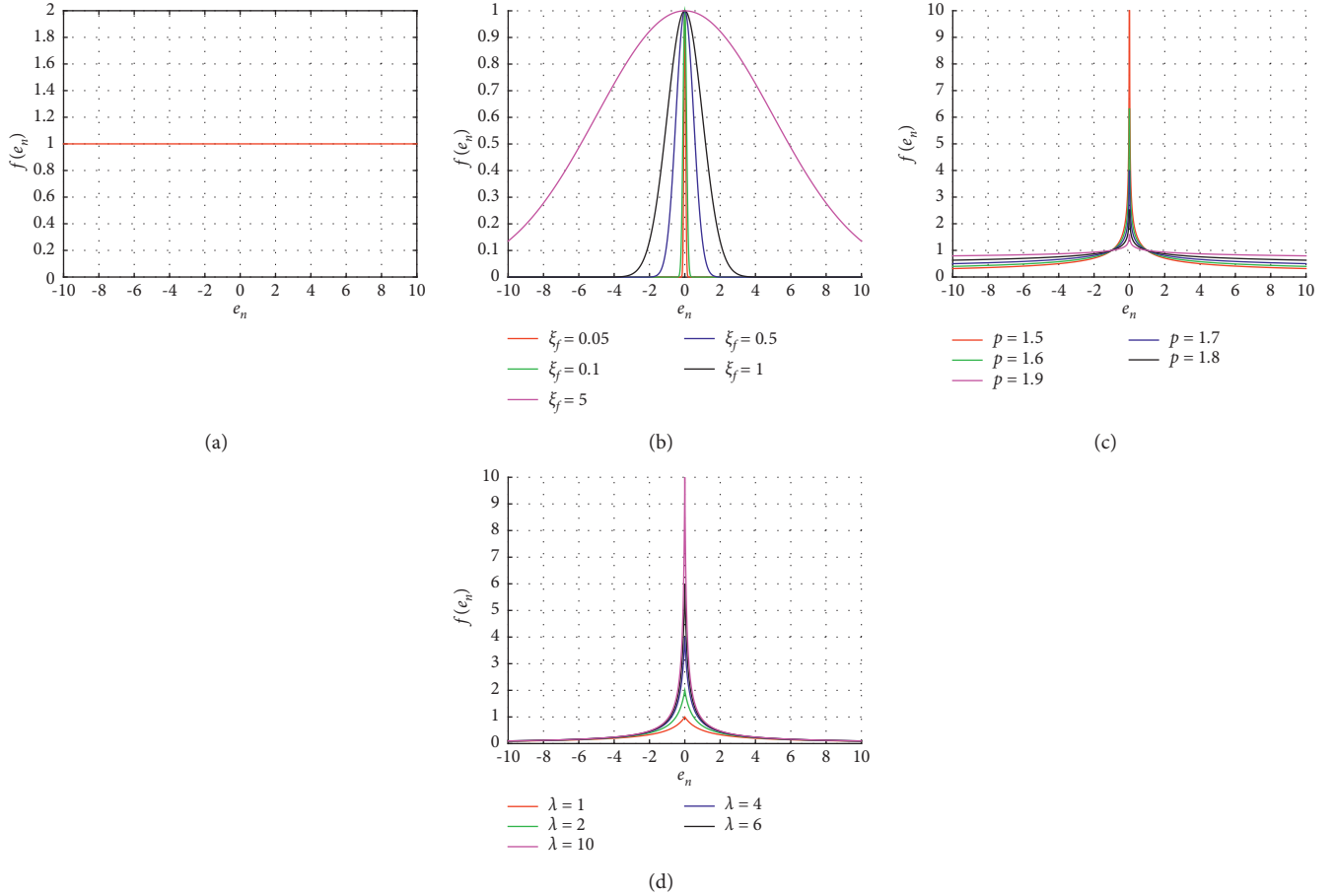


FIGURE 1: Curves of error functions $f(e_n)$ with $e_n \in [-10, 10]$. (a) KLMs algorithm. (b) KMCC algorithm. (c) KLMP algorithm. (d) KLLAD algorithm.

$$\mathbb{E} \left\{ \frac{\tilde{\mathbf{k}}_{\omega,n} \tilde{\mathbf{k}}_{\omega,n}^T}{1 + \lambda |e_n|} \right\} \approx \frac{\mathbb{E} \left\{ \tilde{\mathbf{k}}_{\omega,n} \tilde{\mathbf{k}}_{\omega,n}^T \right\}}{1 + \lambda \mathbb{E} \left\{ |e_n| \right\}}. \quad (30)$$

On the one hand, the numerator of (30) is the autocorrelation matrix of the kernelized input vector defined by

$$\mathbf{R}_{\kappa\kappa} = \mathbb{E} \left\{ \tilde{\mathbf{k}}_{\omega,n} \tilde{\mathbf{k}}_{\omega,n}^T \right\} \in \mathbb{R}^{M \times M}. \quad (31)$$

It has been determined in the theoretical analysis of KAF algorithms [34, 41]. The (i, j) -th element of matrix $\mathbf{R}_{\kappa\kappa}$ can be computed by

$$\begin{aligned} [\mathbf{R}_{\kappa\kappa}]_{ij} &= \left| \mathbf{I}_L + \frac{2}{\xi^2} \mathbf{R}_{xx} \right|^{-1/2} \\ &\times \exp \left(-\frac{1}{4\xi^2} \left[2 \left(\|\mathbf{x}_{\omega,i}\|_2^2 + \|\mathbf{x}_{\omega,j}\|_2^2 \right) \right. \right. \\ &\left. \left. - \|\mathbf{x}_{\omega,i} + \mathbf{x}_{\omega,j}\|_2^2 \left(\mathbf{I}_L + \xi^2 \mathbf{R}_{xx}^{-1/2} \right)^{-1} \right] \right), \end{aligned} \quad (32)$$

with the input covariance matrix $\mathbf{R}_{xx} = \mathbb{E} \{ \mathbf{x}_n \mathbf{x}_n^T \} \in \mathbb{R}^{L \times L}$. On the other hand, using (25) and the MIA assumption, the expression of MSE for KLLAD algorithm is given by

$$\mathbb{E} \{ e_n^2 \} \approx \sigma_z^2 + \text{tr} \{ \mathbf{R}_{\kappa\kappa} \mathbf{V}_n \}, \quad (33)$$

with the autocorrelation matrix of weight error vector $\mathbf{V}_n = \mathbb{E} \{ \mathbf{v}_n \mathbf{v}_n^T \} \in \mathbb{R}^{M \times M}$. Although the recursion of \mathbf{V}_n is not explicitly provided, it has trivial influence on the convergence condition on the mean stability of KLLAD algorithm as shown below. By the result of (4), (33) can be expressed as

$$\mathbb{E} \{ e_n^2 \} = (1 - p_r) \sigma_v^2 + p_r (K + 1) \sigma_v^2 + \text{tr} \{ \mathbf{R}_{\kappa\kappa} \mathbf{V}_n \}. \quad (34)$$

The expectation in denominator of (30) can be roughly approximated as

$$\mathbb{E} \{ |e_n| \} \approx \sqrt{(1 - p_r) \sigma_v^2 + p_r (K + 1) \sigma_v^2 + \text{tr} \{ \mathbf{R}_{\kappa\kappa} \mathbf{V}_n \}}. \quad (35)$$

$$\mathbb{E}\{\mathbf{v}_{n+1}\} = \left(\mathbf{I}_M - \eta\lambda \frac{\mathbf{R}_{\kappa\kappa}}{1 + \lambda\sqrt{(1-p_r)\sigma_v^2 + p_r(K+1)\sigma_v^2 + \text{tr}\{\mathbf{R}_{\kappa\kappa}\mathbf{V}_n\}}} \right) \mathbb{E}\{\mathbf{v}_n\}. \quad (36)$$

Then, (29) can be further determined by (36), as shown at the top of next page. By (36), thus the sufficient convergence condition on mean stability is given by

$$0 < \eta < \frac{2 + 2\lambda\sqrt{(1-p_r)\sigma_v^2 + p_r(K+1)\sigma_v^2 + \text{tr}\{\mathbf{R}_{\kappa\kappa}\mathbf{V}_n\}}}{\text{eig}_{\max}\{\mathbf{R}_{\kappa\kappa}\}}. \quad (37)$$

Thus, (37) can be rigorously reformulated as

$$0 < \eta < \frac{2 + 2\lambda\sigma_v\sqrt{1 + p_rK}}{\text{eig}_{\max}\{\mathbf{R}_{\kappa\kappa}\}}, \quad (38)$$

where the calculation of autocorrelation matrix of weight error vector \mathbf{V}_n is not used as explained before.

5. Simulation Results

In this section, we evaluated the performance of the proposed KLLAD algorithm in the context of impulsive noise by the simulation results. All the curves are obtained by averaging over 200 independent Monte Carlo trails.

5.1. Example I. In general, the optimal weight vectors of KAF algorithms, which correspond to the elements of online dictionary built by an online manner, are unavailable a priori. As a consequence, it is difficult to exhibit the convergence performance, particularly the mean-square-deviation (MSD). In order to obviously demonstrate the excellent mean and mean square convergence performance of KLLAD algorithm, the desired output of an ideal nonlinear synthesis system consisting of the optimum weight vector and the kernelized inputs is given by

$$y_n = \sum_{m=1}^M w_m^* \exp\left(-\frac{\|\mathbf{x}_{\omega,m} - \mathbf{x}_n\|_2^2}{2\xi^2}\right) + z_n, \quad (39)$$

where the kernel bandwidth is set to $\xi = 0.65$, and z_n is the non-Gaussian impulsive CG noise with parameters $K = 1 \times 10^4$, $\sigma_v^2 = 1$, and $p_r = 0.05$.

Furthermore, the CG noise model defined by (2) can be generalized into the mixture Gaussian noise model by changing the distribution of the random variable ϵ_n into the uniform, binary, Rayleigh, Laplacian distributions, etc. Then, we are able to thoroughly investigate the variations of nonlinear filtering performance of KLLAD algorithm in the presence of distinct impulsive noises with large outliers.

Meanwhile, the preselected dictionary with 5 elements utilized for the kernelized nonlinear mapping is given by

$$\begin{aligned} \mathcal{D} &= \{\mathbf{x}_{\omega,1}, \mathbf{x}_{\omega,2}, \mathbf{x}_{\omega,3}, \mathbf{x}_{\omega,4}, \mathbf{x}_{\omega,5}\} \\ &= \left\{ \begin{bmatrix} 0.72 \\ 1.44 \end{bmatrix}, \begin{bmatrix} 3.31 \\ 1.28 \end{bmatrix}, \begin{bmatrix} -3.03 \\ -2.75 \end{bmatrix}, \begin{bmatrix} 1.48 \\ -1.66 \end{bmatrix}, \begin{bmatrix} -1.28 \\ -0.32 \end{bmatrix} \right\}, \end{aligned} \quad (40)$$

which is generated by the coherence criterion from the input signals a priori. Correspondingly, the learning curve of the MSD is defined as

$$\text{MSD}_n = \frac{1}{T} \sum_{t=1}^T \|\mathbf{w}_{t,n} - \mathbf{w}^*\|_2^2, \quad (41)$$

where T is the total number of Monte Carlo runs. The input signal was assumed to be a sequence of statistically independent vectors $\mathbf{x}_n = [x_{1,n}, x_{2,n}]^T$ with correlated samples satisfying $x_{1,n} = 0.5x_{2,n} + \tau_{x,n}$, where $x_{2,n}$ is a white Gaussian noise sequence with variance $\sigma_{x_2}^2 = 1$ and $\tau_{x,n}$ is a white Gaussian so that $x_{1,n}$ has variance $\sigma_{x_1}^2 = 1$. The optimum weight vector \mathbf{w}^* was supposed to be abruptly changed from $[0.3, 0.1, -0.15, -0.3, -0.1]^T$ to $[0.05, 0.25, -0.05, -0.2, -0.35]^T$ at time instant $n = 2 \times 10^4$.

The set of parameters of all used algorithms for Example I is listed in Table 2. It should be pointed that the step-size of KLLAD algorithm is selected by satisfying the convergence condition presented by (38).

The convergence curves of weight coefficients obtained by implementing KLMS, KMCC, KLMP, and KLLAD algorithms are illustrated in Figures 2–5, respectively. As illustrated in Figure 2, the averaged curves of weight coefficients of KLMS algorithm are not able to tend to the optimal weight coefficients during the two stages due to the interference effect of impulsive CG noise. In contrast, the averaged curves of weight coefficients of KMCC, KLMP, and KLLAD algorithms are smooth and all converge to two optimal weight coefficients within the two stages, as shown in Figures 3–5. In addition, the effectiveness of convergence condition on the mean stability (38) is validated by Figure 5. More importantly, Figures 6(a)–6(f) show that the filtering performance of KLLAD algorithm consistently outperforms those of KLMS, KMCC, and KLMP algorithms in terms of robustness, convergence rate, and accuracy of the learning curves of MSD during the two stages using the mixture Gaussian noise model based on the normal, uniform, binary, Rayleigh, Laplacian, and symmetric α -stable distributions, respectively. Therefore, the robust performance of KLLAD algorithm is validated by the simulation results of nonstationary nonlinear system identification in the presence of non-Gaussian impulsive noise.

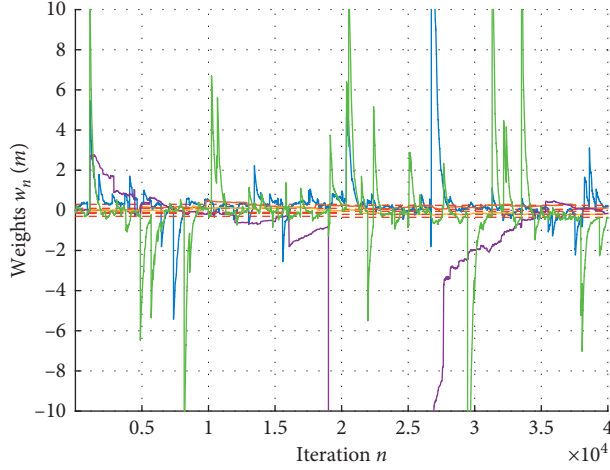
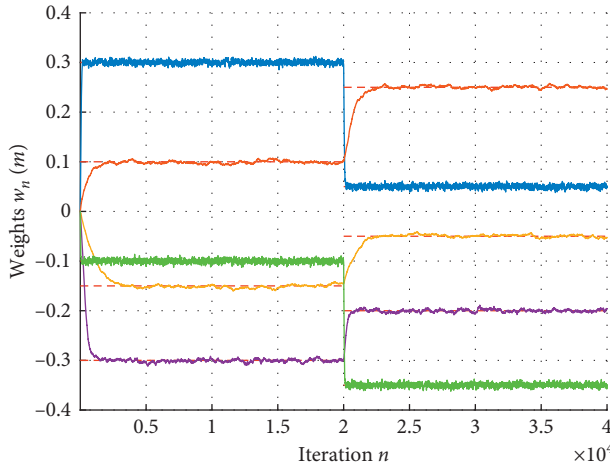
5.2. Example II. As the second example, we consider the input random sequence generated from the following relation:

$$u_n = \rho u_{n-1} + \sigma_u \sqrt{1 - \rho^2} \zeta_n, \quad (42)$$

where ζ_n is a random noise following the i.i.d. standard normal distribution. Here, the correlation factor ρ and the standard deviation σ_u of random sequence u_n were all

TABLE 2: Parameter settings of used algorithms for Example I.

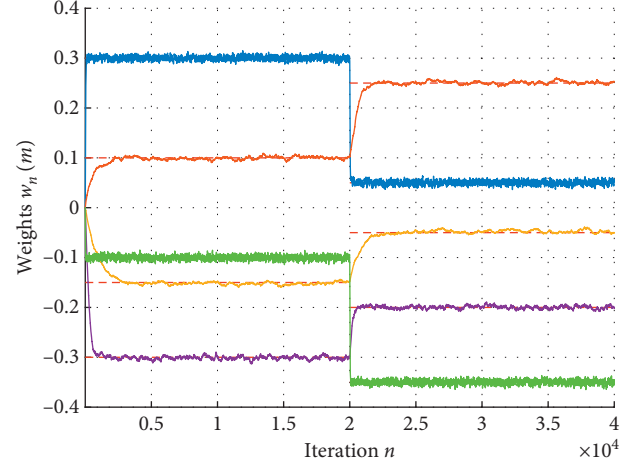
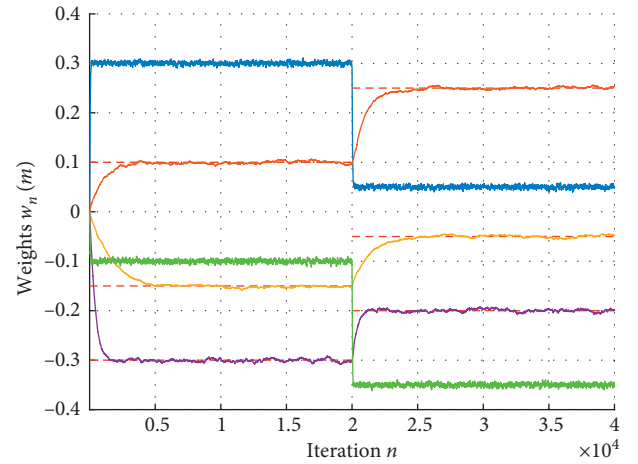
Algorithm	Parameters
KLMS	$\xi = 0.65; \eta = 0.05;$
KMCC	$\xi = 0.65; \eta = 3; \xi_f = 0.05;$
KLMP	$\xi = 0.65; \eta = 0.05; p = 1.1$
KLLAD	$\xi = 0.65; \eta = 0.08; \lambda = 15$

FIGURE 2: Convergence of the coefficients $w_n(m)$ for KLMS algorithm (the dotted lines are for the optimal weight coefficients, and solid lines are for weight coefficients of KLMS algorithm).FIGURE 3: Convergence of the coefficients $w_n(m)$ for KMCC algorithm (the dotted lines are for the optimal weight coefficients, and solid lines are for weight coefficients of KMCC algorithm).

chosen as 0.5. The desired output of nonlinear system is generated as follows:

$$\begin{cases} t_n = 0.5u_n - 0.3u_{n-1}, \\ y_n = t_n - 1.25t_n^2 + 0.25t_n^3 + z_n, \end{cases} \quad (43)$$

where z_n is the non-Gaussian impulsive CG noise with parameters $K = 1 \times 10^4$, $\sigma_v^2 = 1$, and $p_r = 0.1$. At each time instant n , the input vectors $\mathbf{x}_n = [u_n, u_{n-1}]^T$ are used to estimate the nonlinear desired response signal y_n contaminated by the impulsive noise. In order to clearly investigate the characteristic of convergence of weight coefficients, the 6

FIGURE 4: Convergence of the coefficients $w_n(m)$ for KLMP algorithm (the dotted lines are for the optimal weight coefficients, and solid lines are for weight coefficients of KLMP algorithm).FIGURE 5: Convergence of the coefficients $w_n(m)$ for KLLAD algorithm (the dotted lines are for the optimal weight coefficients, and solid lines are for weight coefficients of KLLAD algorithm).

elements of dictionary were chosen by the coherence criterion from the input vectors a priori as follows:

$$\begin{aligned} \mathcal{D} &= \{\mathbf{x}_{\omega,1}, \mathbf{x}_{\omega,2}, \mathbf{x}_{\omega,3}, \mathbf{x}_{\omega,4}, \mathbf{x}_{\omega,5}, \mathbf{x}_{\omega,6}\} \\ &= \left\{ \begin{bmatrix} 0.17 \\ -1.92 \end{bmatrix}, \begin{bmatrix} -1.62 \\ -0.18 \end{bmatrix}, \begin{bmatrix} 0.52 \\ 1.55 \end{bmatrix}, \begin{bmatrix} 2.90 \\ 1.92 \end{bmatrix}, \begin{bmatrix} -2.0 \\ -2.47 \end{bmatrix}, \begin{bmatrix} 2.65 \\ -0.82 \end{bmatrix} \right\}. \end{aligned} \quad (44)$$

The set of parameters of used algorithms for Example II is listed in Table 3. Likewise, the step-size of KLLAD algorithm in second example is also determined from the range of convergence condition according to (38).

Although the fixed dictionary with 6 elements is used to compare the differences of convergent performance among used algorithms, the corresponding optimal weight

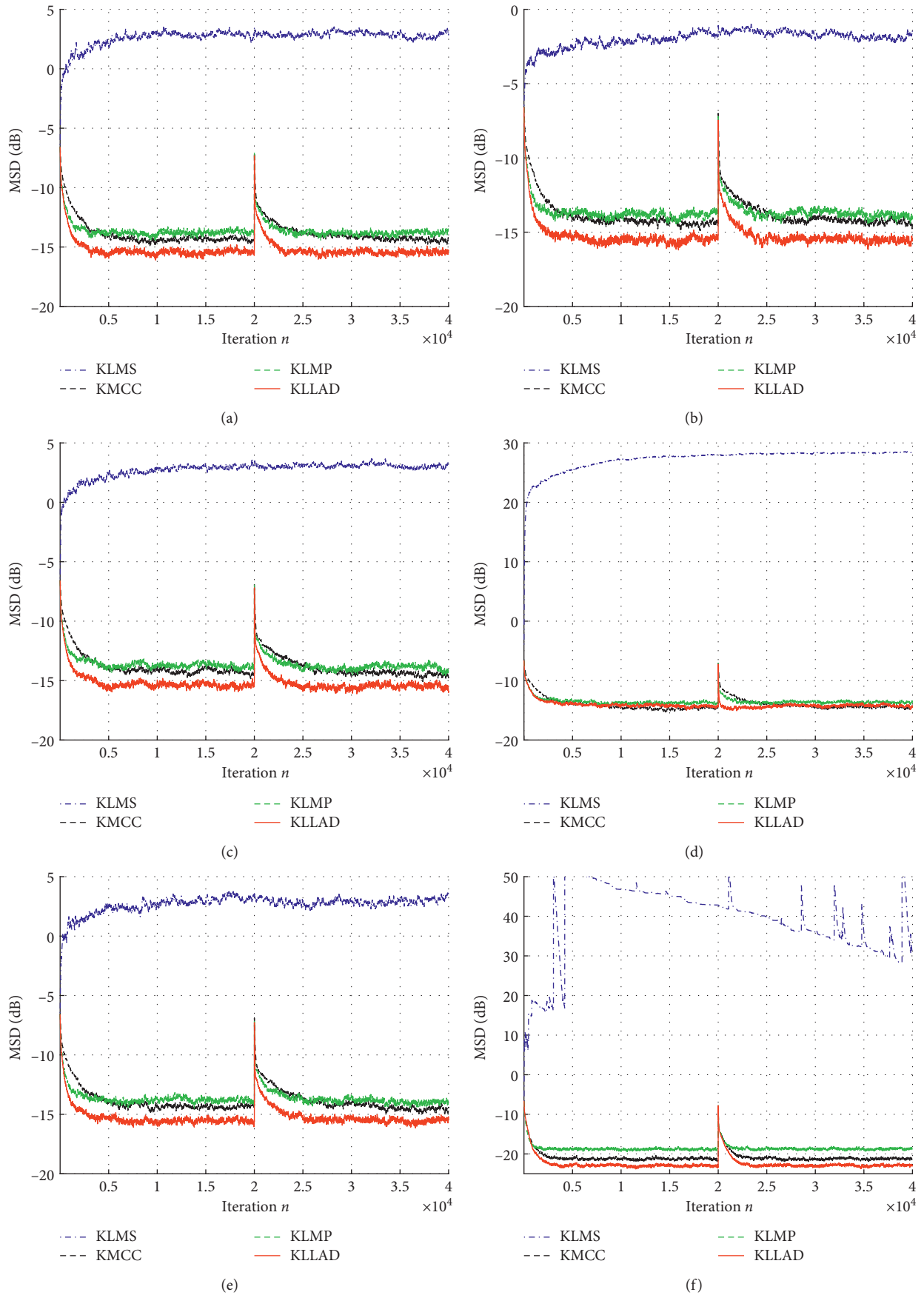
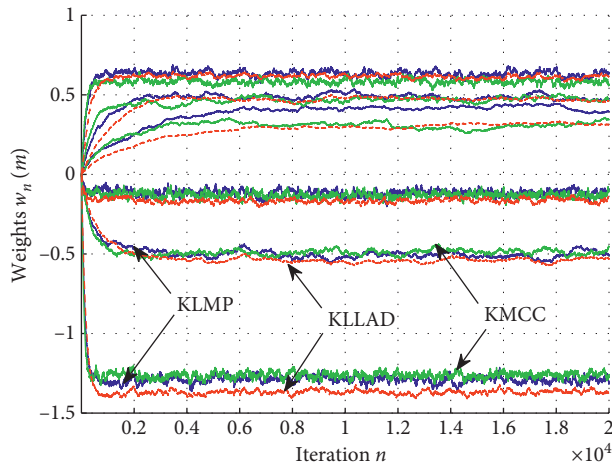


FIGURE 6: Comparisons of learning curves of MSD for KLMS, KMCC, KLMP, and KLLAD algorithms with different non-Gaussian distribution noises. (a) Normal distribution noise. (b) Uniform distribution noise with $(-1, 1)$. (c) Binary distribution noise with $(-1, 1)$. (d) Rayleigh distribution noise with $\sigma^2 = 4$. (e) Laplacian distribution noise with $\mu = 0$ and $\sigma^2 = 1$. (f) Symmetric α -stable distribution noise with $\alpha = 0.9$, $\gamma = 0.5$, and $\delta = 0$.

TABLE 3: Parameter settings of used algorithms for Example II.

Algorithm	Parameters
KMCC	$\xi = 0.75; \eta = 1; \xi_f = 0.5;$
KLMP	$\xi = 0.75; \eta = 0.1; p = 0.7$
KLLAD	$\xi = 0.75; \eta = 0.1; \lambda = 15$

FIGURE 7: Convergence of the coefficients $w_n(m)$ for KMCC, KLMP, and KLLAD algorithms.

coefficients are still unknown in fact. The KLMS algorithm is ignored due to its severe instability of convergence. Figure 7 shows that the mean value curves of weight coefficients of KLLAD algorithm converge much smoother and faster compared with the KLMP and KMCC algorithms, which verified the superiorities of KLLAD algorithm.

6. Conclusion

In this paper, we presented a novel KLLAD algorithm based on the logarithmic error cost criterion under the assumption of non-Gaussian impulsive CG noise. The KLLAD algorithm can effectively mitigate the instability of convergence learning curves caused by impulsive noise. The simulation results demonstrated that the proposed KLLAD algorithm has an excellent performance compared to the KLMS, KMCC, and KLMP algorithms in the presence of impulsive noise.

Data Availability

The ideal mathematical model data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This research was supported by the National Natural Science Foundation of China under Grant nos. 61701200 and 61701529.

References

- [1] C. Richard, J. C. M. Bermudez, and P. Honeine, "Online prediction of time series data with kernels," *IEEE Transactions on Signal Processing*, vol. 57, no. 3, pp. 1058–1067, 2009.
- [2] J. M. Gil-Cacho, M. Signoretto, T. van Waterschoot, M. Moonen, and S. H. Jensen, "Nonlinear acoustic echo cancellation based on a sliding-window leaky kernel affine projection algorithm," *IEEE Transactions on Audio Speech and Language Processing*, vol. 21, no. 9, pp. 1867–1878, 2013.
- [3] Y. Nakajima and M. Yukawa, "Nonlinear channel equalization by multi-kernel adaptive filter," in *Proceedings of the IEEE 13th International Workshop on Signal Processing Advances in Wireless Communications*, pp. 384–388, SPAWC, Cesme, Turkey, June, 2012.
- [4] T. Wang, J. Chen, P. Honeine, and H. Snoussi, "Abnormal event detection via multikernel learning for distributed camera networks," *International Journal of Distributed Sensor Networks*, vol. 2015, Article ID 989450, 213 pages, 2015.
- [5] W. Liu, J. C. Principe, and S. Haykin, *Kernel Adaptive Filtering: A Comprehensive Introduction*, John Wiley & Sons, New-York, NY, USA, 2010.
- [6] M. Yukawa, "Multikernel adaptive filtering," *IEEE Transactions on Signal Processing*, vol. 60, no. 9, pp. 4672–4682, 2012.
- [7] B. Chen, S. Zhao, P. Zhu, and J. C. Principe, "Quantized kernel recursive least squares algorithm," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 9, pp. 1484–1491, 2013.
- [8] M. Zhang, X. Wang, X. Chen, and A. Zhang, "The kernel conjugate gradient algorithms," *IEEE Transactions on Signal Processing*, vol. 66, no. 16, pp. 4377–4387, 2018.
- [9] W. Liu, P. P. Pokharel, and J. C. Principe, "The kernel least-mean-square algorithm," *IEEE Transactions on Signal Processing*, vol. 56, no. 2, pp. 543–554, 2008.
- [10] M. A. Chitre, J. R. Potter, and S.-H. Ong, "Optimal and near-optimal signal detection in snapping shrimp dominated ambient noise," *IEEE Journal of Oceanic Engineering*, vol. 31, no. 2, pp. 497–503, 2006.
- [11] A. Mahmood, M. Chitre, and M. A. Armand, "Detecting OFDM signals in alpha-stable noise," *IEEE Transactions on Communications*, vol. 62, no. 10, pp. 3571–3583, 2014.
- [12] J. Ilow and D. Hatzinakos, "Analytic alpha-stable noise modeling in a Poisson field of interferers or scatterers," *IEEE Transactions on Signal Processing*, vol. 46, no. 6, pp. 1601–1611, 1998.
- [13] J. Ni, J. Chen, and X. Chen, "Diffusion sign-error LMS algorithm: formulation and stochastic behavior analysis," *Signal Processing*, vol. 128, pp. 142–149, 2016.
- [14] W. Gao, Y. Xu, and J. Chen, "Stochastic analysis of diffusion dual sign algorithm for cyclostationary white Gaussian inputs and nonstationary system," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 67, no. 7, pp. 1354–1358, Jul. 2020.
- [15] W. Gao and J. Chen, "Transient analysis of signed LMS algorithms with cyclostationary colored Gaussian inputs," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 67, no. 12, pp. 3562–3566, 2020.
- [16] B. Chen, L. Xing, H. Zhao, N. Zheng, and J. C. Principe, "Generalized correntropy for Robust_newline adaptive filtering," *IEEE Transactions on Signal Processing*, vol. 64, no. 13, pp. 3376–3387, 2016.
- [17] W. Ma, B. Chen, H. Zhao, G. Gui, J. Duan, and J. C. Principe, "Sparse least logarithmic absolute difference algorithm with correntropy-induced metric penalty," *Circuits, Systems, and Signal Processing*, vol. 35, no. 3, pp. 1077–1089, 2016.

- [18] B. Chen, L. Xing, B. Xu, H. Zhao, N. Zheng, and J. C. Principe, "Kernel risk-sensitive loss: definition, properties and application to robust adaptive filtering," *IEEE Transactions on Signal Processing*, vol. 65, no. 11, pp. 2888–2901, 2017.
- [19] V. C. Gogineni and S. Mula, "Logarithmic cost based constrained adaptive filtering algorithms for sensor array beamforming," *IEEE Sensors Journal*, vol. 18, no. 14, pp. 5897–5905, 2018.
- [20] K. Xiong and S. Wang, "Robust least mean logarithmic square adaptive filtering algorithms," *Journal of the Franklin Institute*, vol. 356, no. 1, pp. 654–674, 2019.
- [21] B. Binwei Weng and K. E. Barner, "Nonlinear system identification in impulsive environments," *IEEE Transactions on Signal Processing*, vol. 53, no. 7, pp. 2588–2594, 2005.
- [22] S. Zhao, B. Chen, and J. C. Principe, "Kernel adaptive filtering with maximum correntropy criterion," in *Proceedings of the The 2011 International Joint Conference on Neural Networks*, pp. 2012–2017, San Jose, CA, USA, August, 2011.
- [23] S. Wang, J. Feng, and C. K. Tse, "Kernel affine projection sign algorithms for combating impulse interference," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 60, no. 11, pp. 811–815, 2013.
- [24] L. Lu, H. Zhao, and B. Chen, "Improved-variable-forgetting-factor recursive algorithm based on the logarithmic cost for Volterra system identification," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 63, no. 6, pp. 588–592, 2016.
- [25] Y. He, F. Wang, J. Yang, H. Rong, and B. Chen, "Kernel adaptive filtering under generalized maximum correntropy criterion," in *Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN)*, pp. 1738–1745, Vancouver, BC, Canada, July, 2016.
- [26] X. Luo, J. Deng, W. Wang, J.-H. Wang, and W. Zhao, "A quantized kernel learning algorithm using a minimum kernel risk-sensitive loss criterion and bilateral gradient technique," *Entropy*, vol. 19, no. 7, p. 365, 2017.
- [27] X. Li, Q. Lu, Y. Dong, and D. Tao, "Robust subspace clustering by Cauchy loss function," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 7, pp. 2067–2078, 2019.
- [28] N. Guan, T. Liu, Y. Zhang, D. Tao, and L. S. Davis, "Truncated Cauchy non-negative matrix factorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 246–259, 2019.
- [29] W. Shi, K. Xiong, and S. Wang, "Multikernel adaptive filters under the minimum Cauchy kernel loss criterion," *IEEE Access*, vol. 7, pp. 120 548–120 558, 2019.
- [30] Q. Wu, Y. Li, Y. V. Zakharov, W. Xue, and W. Shi, "A kernel affine projection-like algorithm in reproducing kernel Hilbert space," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 67, no. 10, pp. 2249–2253, 2020.
- [31] W. Gao and J. Chen, "Kernel least mean p -Power algorithm," *IEEE Signal Processing Letters*, vol. 24, no. 7, pp. 996–1000, Jul. 2017.
- [32] W. Ma, J. Duan, W. Man, H. Zhao, and B. Chen, "Robust kernel adaptive filters based on mean p -power error for noisy chaotic time series prediction," *Engineering Applications of Artificial Intelligence*, vol. 58, pp. 101–110, 2017.
- [33] T. Zhang and S. Wang, "Nyström kernel algorithm under generalized maximum correntropy criterion," *IEEE Signal Processing Letters*, vol. 27, pp. 1535–1539, 2020.
- [34] W. Gao, M. Song, and J. Chen, "Tracking analysis of Gaussian kernel signed error algorithm for time-variant nonlinear systems," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 67, no. 10, pp. 2289–2293, 2020.
- [35] S. Wang, W. Wang, K. Xiong, H. H. C. Iu, and C. K. Tse, "Logarithmic hyperbolic Cosine adaptive filter and its performance analysis," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 51, no. 4, pp. 2512–2524, 2021.
- [36] H. Zhang, B. Yang, L. Wang, and S. Wang, "General Cauchy conjugate gradient algorithms based on multiple random Fourier features," *IEEE Transactions on Signal Processing*, vol. 69, pp. 1859–1873, 2021.
- [37] M. O. Sayin, N. D. Vanli, and S. S. Kozat, "A novel family of adaptive filtering algorithms based on the logarithmic cost," *IEEE Transactions on Signal Processing*, vol. 62, no. 17, pp. 4411–4424, 2014.
- [38] Y. Zou, S. C. Chan, and T. S. Ng, "A recursive least m-estimate (RLM) adaptive filter for robust filtering in impulse noise," *IEEE Signal Processing Letters*, vol. 7, no. 11, pp. 324–326, 2000.
- [39] B. Schölkopf, R. Herbrich, and R. Williamson, *A generalized representer theorem*, Vol. 81, Springer, , Berlin, Heidelberg, 2000.
- [40] Y. Engel, S. Mannor, and R. Meir, "The kernel recursive least-squares algorithm," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2275–2285, 2004.
- [41] W. Gao and J. Chen, "Transient performance analysis of zero-attracting Gaussian kernel LMS algorithm with pre-tuned dictionary," *IEEE Access*, vol. 7, pp. 135 770–135 779, 2019.
- [42] W. Gao, J. Chen, C. Richard, and J. Huang, "Online dictionary learning for kernel LMS," *IEEE Transactions on Signal Processing*, vol. 62, no. 11, pp. 2765–2777, 2014.

Research Article

Radar-Based Rainfall Estimation of Landfalling Tropical Storm “PABUK” 2019 over Southern Thailand

Pakdee Chantraket ^{1,2}, **Sukrit Kirtsaeng** ³, **Chakrit Chaotamonsak**^{4,5},
Somporn Chantara ⁴, **Supachai Nakapan**⁴ and **Thammarat Panityakul** ⁶

¹Department of Royal Rainmaking and Agricultural Aviation, Bangkok 10900, Thailand

²PhD Degree Program in Environmental Science, Environmental Science Research Center, Faculty of Science, Chiang Mai University, Chiang Mai 50200, Thailand

³Thai Meteorological Department, Bangkok 10260, Thailand

⁴Regional Centers for Climate and Environmental Studies (RCCES), Chiang Mai University, Chiang Mai 50200, Thailand

⁵Department of Geography, Faculty of Social Sciences, Chiang Mai University, Chiang Mai 50200, Thailand

⁶Division of Computational Science, Faculty of Science, Prince of Songkla University, Songkhla 90110, Thailand

Correspondence should be addressed to Pakdee Chantraket; pakdee2@gmail.com

Received 9 March 2021; Revised 3 November 2021; Accepted 15 November 2021; Published 2 February 2022

Academic Editor: Harish Garg

Copyright © 2022 Pakdee Chantraket et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Tropical storm PABUK developed from tropical depression first defined on 31 December 2018 in the lower South China Sea. It made landfall in Pak Phanang, Nakhon Si Thammarat province, southern Thailand. PABUK caused heavy rain and flash floods from 3 to 5 January 2019 (D1, D2, and D3) where the total rainfall reached 150–300 mm across 14 provinces of southern Thailand. This paper is aimed to investigate rainstorm properties and rainfall estimation of tropical storm PABUK with weather radar in southern Thailand. The radar data analysis in this study was to extract the radar reflectivity to study rainstorm properties of PABUK over 3 days along southern Thailand derived from the Thunderstorm Identification and Tracking Analysis and Nowcasting (TITAN) algorithm including 5 variables of duration, area, cloud-based height, maximum reflectivity, and speed in the data set. Based on the properties and frequency distribution of 2,557 rainstorms in D1, D2, and D3, rainstorms in D2 and D3 when PABUK made landfall over southern Thailand show a longer lifetime, higher reflectivity, and larger rain-cells as well as it was found efficient in terms of rainfall amount than in D1. In addition, the estimated rainfall using weather radar provides important information of the rainfall distribution for the analysis of the rainstorm as well. These analyses provide a context for interpreting the feasible rainfall estimates based on Z-R relationship during tropical storm PABUK that produced extreme floods in southern Thailand. A Z-R relationship in the form $Z = 104R^{1.5}$ provided acceptable statistical indicators, making it appropriate for radar estimated rainfall in case studies presented of tropical storm PABUK in southern Thailand. However, the result of this study should be improved to estimate precipitation in case of extremely heavy rainfall in tropical storm occurrence by using radar of southern Thailand and applied for applications of early warning systems.

1. Introduction

A large number of studies on rainstorm analysis obtained from capability for spatial and temporal storm profiles measurement of weather radar are widely used to detect the convective storms and study of convective storm structure [1–3] and also the process of the rainfall system itself by providing real-time regional information, and with the

existence of long radar data sets, these data could be also applied for climatological applications. In addition, weather radar is also tool that combines meteorology and hydrology [4–6]; the meteorological information measured by radar are used for hydrological analysis as referred to Peng et al. [7], and they explained that the advantage of using radar for precipitation measurement is the coverage of a large area in real-time, and radars also experience difficulty in achieving

an accurate estimation for hydrological applications. The single-polarization [3] and dual-polarization [8, 9] weather radar was used to study the relationship between radar reflectivity and rainfall rate which is developed for rainfall measurement by using the Z-R relationship. The uncertainty of radar rainfall estimation by using Z-R relationship was proposed by Chen et al. [10] and Gou et al. [11]. The developed Z-R relationships are needed in Thailand to provide a more systematic and comprehensive approach to achieve water management.

Applications of weather radar in Thailand are still limited mostly for meteorology and monitoring the weather routines. Not much work has been done in the field of hydrological and heavy rainfall cases. Finding rainfall intensity is one of the essential applications for weather radar in the process of hydrology, flood management, and early warning system in case of severe weather situation. For the purpose in radar rainfall estimation, the relationship between radar reflectivity and rainfall rate is developed for rainfall measurement by using the Z-R relationship. The Z-R relationship is highly dependent on the precipitation types and wind conditions such as convective, stratiform or mixed types, and deep convection [12–15]. Event type is one of the major influences of Z-R relationship that must be studied accordingly. Moreover, the location of areas and seasonal also plays an important factor in applying Z-R relationships to radar rainfall measurements [1, 16]. Most weather radars in Thailand are not calibrated for the Z-R relationship. As the results, the developed radar rainfall estimation in case of severe weather situation from tropical storm is needed in Thailand providing a more systematic and comprehensive approach to achieve in water management and also additionally to implement in flood warning purposes. Because the southern part of Thailand is a major economic tree plantation zone, especially fruits, oil palm, and rubber tree, it is almost in transition from water richness to water scarcity because of the increasing demands on this limited resource as well as there is no universal Z-R relationship that can be applied to all cases of rainfall events. Therefore, the focus of this paper will be on the optimization of Z-R relationships during tropical storm and heavy precipitation which were tuned to fit the rain gauge measurements that turn into inaccuracies over the central region of Thailand.

This article is structured as follows. The data and methodology section presents the overview of PABUK tropical cyclone, the technical characteristics of the radar used, radar data analyses, and the statistics for the analysis. This is followed with presenting the results of rainstorm properties and radar rainfall estimation during PABUK event, and the article closes with a brief discussion about the conclusion reached.

2. Data and Methodology

2.1. Overview of PABUK Tropical Cyclone. Tropical storm PABUK, which has originated from the low-pressure zone in the South China Sea, developed from tropical depression first defined on 31 December 2018. This storm moved westward into the lower Gulf of Thailand, and it made

landfall on 4 January 2019 over Phanang, Nakhon Si Thammarat province, at latitude of 8.2°N and longitude of 100.2°E. Maximum sustained wind is 75 kmhr⁻¹, and the storm was moving northwest at a speed of 18 kmhr⁻¹. PABUK became the first tropical storm to make landfall over southern Thailand since Linda in 1997. This affected the South with widespread heavy rainfalls, and torrential downpours are possible in much of southern area of Thailand from 3 to 5 January 2019 as follows:

- (1) On 3 January 2019 (hereafter referred as D1) at 11.00 UTC, tropical storm “PABUK” was located 500 km southeast of Nakhon Si Thammarat province at the latitude of 6.5°N and longitude of 104.2°E with maximum sustained winds of 65 kmhr⁻¹. The storm was accelerated west-northwestward and entered the Gulf of Thailand which affected the lower part of southern Thailand with heavy rainfalls and some torrential downpours in Phatthalung, Songkhla, Pattani, Yala, and Narathiwat.
- (2) On 4 January 2019 (hereafter as D2), the storm moved westward into the lower Gulf of Thailand, and it made landfall at 05:45 UTC over Phanang, Nakhon Si Thammarat province, at latitude of 8.2°N and longitude of 100.2°E with the maximum sustained wind of 75 kmhr⁻¹, and the storm was moving northwest at a speed of 18 kmhr⁻¹. It affected the east side of South with widespread heavy rainfall, strong winds, and severe conditions that cause forest run-offs and flash floods in Phetchaburi, Prachuap Khiri Khan, Chumphon, Surat Thani, Nakhon Si Thammarat, Phatthalung, Ranong, Phangnga, Phuket, Krabi, Trang, and Satun.
- (3) On 5 January 2019 (hereafter as D3) at 17.00 UTC, tropical storm “PABUK” was due 5 km west of Takua Pa, Phangnga, with the latitude of 8.7°N and longitude of 104.2°E with the maximum sustained wind of 55 kmhr⁻¹; the storm was moving west-northwest slowly with the outbreaks of torrential downpours much of southern provinces including Phetchaburi, Prachuap Khiri Khan, Chumphon, Surat Thani, Nakhon Si Thammarat, Ranong, Phangnga, Phuket, and Krabi. PABUK then moved down to the Andaman Sea and weakened into a low-pressure cell during the same day and covered the Andaman Sea.

The influence of tropical storm PABUK caused strong wind shear, heavy rain, and flash floods from D1 to D3 where the accumulated rainfall reached 150–300 mm a day across 14 provinces of southern Thailand, especially in Nakhon Si Thammarat, Surat Thani, Chumphon, Ranong, Phatthalung, Songkhla, Pattani, Yala, and Narathiwat, the maximum rainfall in 24 hours reached 309.3 mm, and maximum wind speed was 89 kmhr⁻¹ at Nakhon Si Thammarat province on D2. In addition, PABUK also results in rising sea levels and blowing into the shore as storm surge in the coastal region of upper southern region; the images of the damage caused by a tropical storm PABUK are as shown in Figure 1.



FIGURE 1: Strong wind shear and flooding caused by tropical Storm PABUK at first hit over Phanang, Nakhon Si ammarat province, southern) ailand on D2 (image: REUTERS and <https://news.mthai.com/>).

2.2. Radar Data Analyses. The study area has been southern Thailand, characterized by a complex topography and directly influenced by the South China Sea and Indian Ocean (Figure 2). The C-Band Doppler Radar, which represents a good compromise between range and reflectivity that can provide rain detection up to a range of 240 km, from Thai Meteorological Department (TMD) was used in this study. The reflectivity data from Songkhla's radar located in the eastern coastline near the landfall of PABUK tropical storm were appropriately used to investigate the rainstorm properties and estimated radar rainfall that affected the southern region of Thailand. Radar is installed at Sathingpra District, Songkhla, at the elevation of 33 m MSL in southern Thailand as shown in Figure 2. The radar with EDGE™ software collected the reflectivity data as volume scan to the highest altitude up to 5 km provided in the universal flies (UF) format [17]. The files were obtained every 15 minutes up to the effective range of 240 km to the highest altitude up to 5 km provided in the volume format files for 4 elevation angles: 0.5°, 1.5°, 2.4°, and 3.4°, and a Doppler filter is applied to remove ground clutter and fixed echoes.

In order to characterize rainstorm properties, radar reflectivity data in horizontal polarization were run through TITAN and used the 30 dBZ reflectivity threshold to identify a convective storm cell before tracking their movement as referred to Dixon and Wiener [18]; Johnson et al. [19], and Potts et al. [20]. All of rainstorms were selected and analyzed by the dataset from the criteria as suggestion by Chantraket et al., [1] including their 5 properties as exhibited in Table 1. However, it may be mentioned that in this study, only those rainstorms are considered whose rain centers are located in the effective range of 240 km of Songkhla's radar. The event numbers differed in each day; these data were then analyzed to express properties of individual rainstorms. The total of 2,557 rainstorm events were chosen from D1 (1,014 rainstorms), D2 (962 rainstorms), and D3 (581 rainstorms), respectively, during occurred PABUK tropical cyclone. The example case of rainstorm events from TITAN analysis is presented in Figure 3, and the preliminary of statistical analysis of all properties is illustrated as Table 2.

In order to estimate radar rainfall, the relationship between radar reflectivity and rainfall rate which is developed for rainfall measurement was used as explained in the following equation:

$$Z = aR^b, \quad (1)$$

where a and b are the relationship parameters, Z is the radar reflectivity in mm^6m^{-3} , and R is the intensity of precipitation mmh^{-1} .

The rainfall events were used to obtain the appropriated Z-R relationship for tropical storm in southern Thailand as well as to test an accuracy of the proposed radar rainfall estimation based on different Z-R relationships as referred to Kirtsang and Chantraket [16] of Songkhla's radar of $Z = 104R^{1.5}$, $Z = 162R^{1.5}$, and $Z = 184R^{1.5}$ and the operational Z-R relationships of TMD $Z = 300R^{1.4}$ as referred for all rain types especially for deep convective [21, 22]. In order to evaluate the suitable Z-R relationship for tropical storm of PABUK, their measurement was compared with the precipitation recorded by the rain gauges from automatic meteorological stations of Hydro Informatics Institute (HII) and TMD. The study was performed by using 24-hour accumulations of 156 rain gauges in and around southern regions in the coverage of radar effective range (see Figure 2) having been scrutinized during D1 to D3. These data were procured from the (i) 129 stations of HII and (ii) 27 stations of TMD. After procurement of 24-hour accumulations from different sources, these were subjected to extensive quality control tests to remove gross errors, archival errors, and reformatting problem; however, the suspected data were checked for validation from different reliable sources. The study was carried out using 24-hour accumulations, comparing with estimated rain accumulation from Z-R relationship that occupied the position of rain gauges. Rain gauge measurement and radar estimate of 24-hour accumulations greater than 1.0 mm and less than 300 mm were considered to be valid. Daily rainfall distribution in southern Thailand caused by tropical storm PABUK during D1, D2, and D3 is shown in Figure 4.

The statistical indexes used to evaluate the different of estimated rain accumulation from Z-R relationship related with the rain gauges as recommended by [23] are as follows.

Root mean square error (RMSE) is as follows:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (R_i - G_i)^2}. \quad (2)$$

Mean error (ME) is as follows:

$$\text{ME} = \frac{1}{n} \sum_{i=1}^N (R_i - G_i). \quad (3)$$

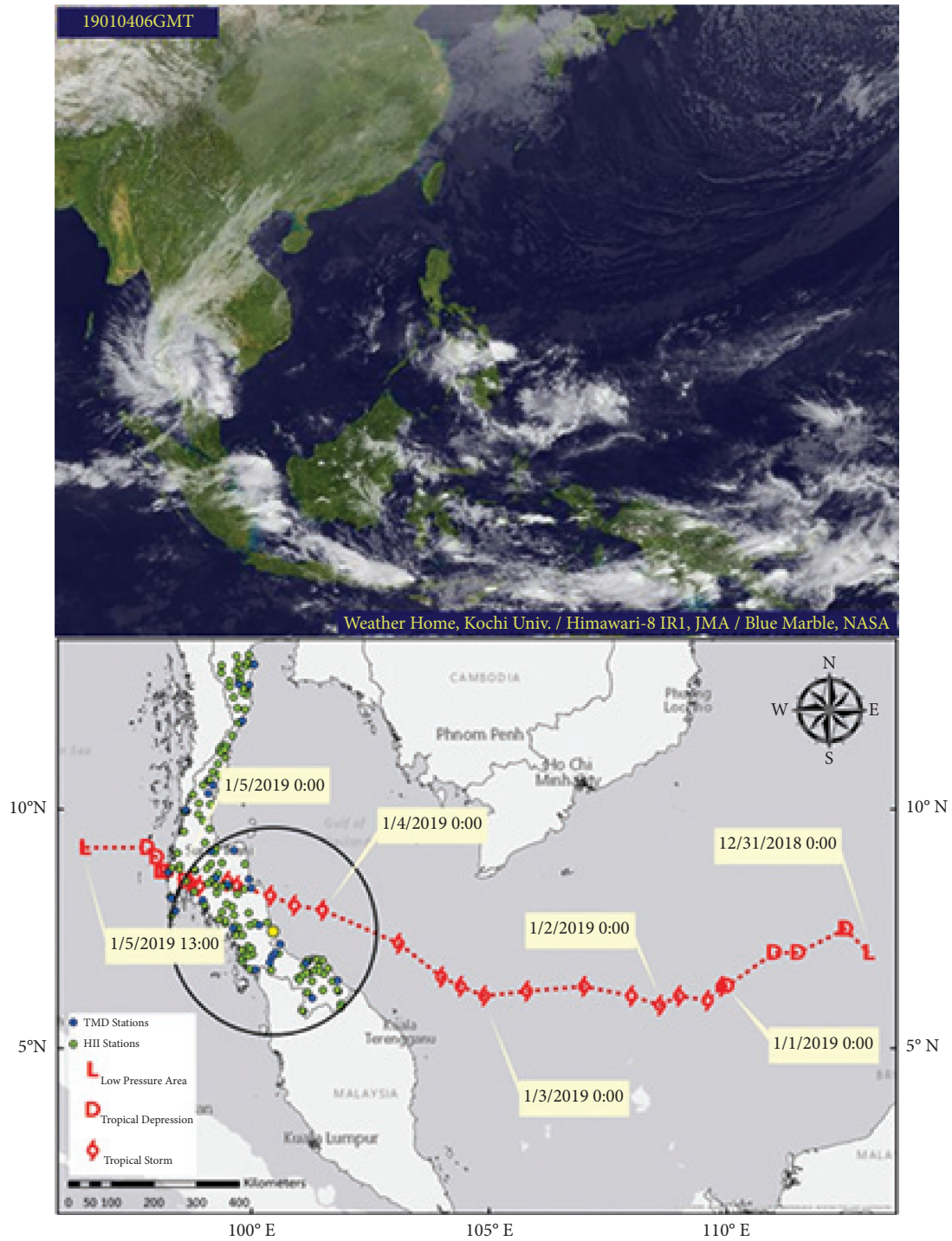
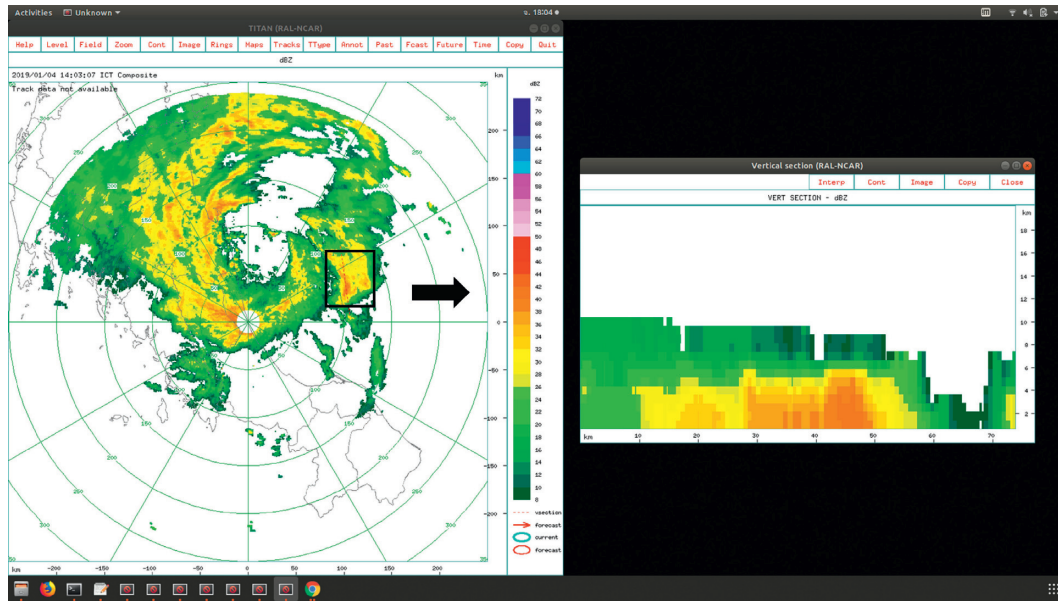


FIGURE 2: Tropical storm PABUK developed from a tropical depression and being placed at 6 h intervals and as observed during D3 at 0600 UTC time by Japan's Himawari-8 satellite. Image credit: JMA (Japan meteorological agency).

TABLE 1: Radar-obtained storm characteristics from TITAN and their units.

Rainstorm properties	Variables	Units
(1) Mean storm duration	SDUR	Hours
(2) Mean storm-based	SBAS	km MSL
(3) Maximum envelope area	SARE	km ²
(4) Maximum reflectivity	SREF	dBZ
(5) Mean speed	SVEL	kmhr ⁻¹



(a)

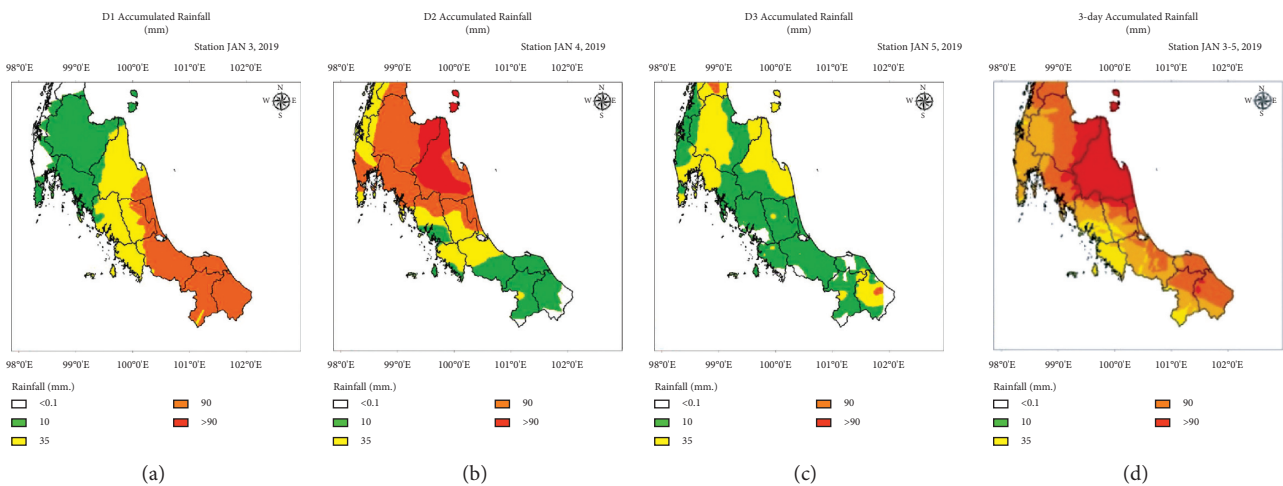
(b)

FIGURE 3: Some case of rainstorm events obtained from TITAN analysis of Songkhla’s radar on D2 at 14:03 UTC: (a) TITAN analysis image and (b) cross section of selected rainstorm.

TABLE 2: The storm properties during D1, D2, and D3 of PABUK event in southern Thailand.

	SDUR				SBAS				SREF				SARE				SVEL			
	Hours				km MSL				dBZ				km ²				kmhr ⁻¹			
	D1	D2	D3	D1-D3	D1	D2	D3	D1-D3	D1	D2	D3	D1-D3	D1	D2	D3	D1-D3	D1	D2	D3	D1-D3
Mean	0.9	0.9	1.0	0.9	2.0	1.8	1.9	1.9	35.0	38.6	43.5	38.3	52.0	55.8	46.4	52.2	23.1	22.5	18.2	21.8
SD	0.4	0.5	0.7	0.5	0.9	0.9	0.9	0.9	5.2	6.9	8.3	7.4	112.9	138.7	68.1	115.5	13.4	13.5	11.5	13.2
Min	0.5	0.5	0.5	0.5	0.6	0.6	0.6	0.6	30.0	30.0	30.0	30.0	3.4	4.5	4.5	3.4	0.0	0.0	0.0	0.0
Max	2.8	4.7	4.8	4.8	3.6	3.6	3.6	3.6	61.0	62.0	67.5	67.5	2666.3	3359.3	737.4	3359.3	62.3	114.5	56.0	114.5

D1: $N=1014$ events, D2: $N=962$ events, and D3: $N=581$ events.



(a)

(b)

(c)

(d)

FIGURE 4: Rainfall distribution of 24-hour accumulations in southern Thailand caused by tropical storm PABUK through (a) D1, (b) D2, and, (c) D3 and (d) 3-day accumulated rainfall during D1 to D3.

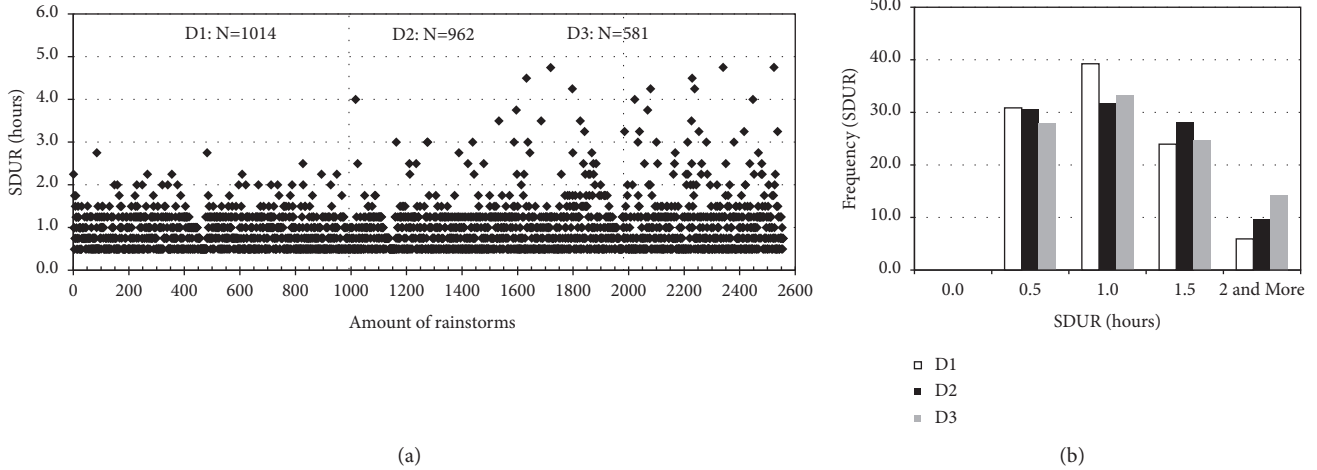


FIGURE 5: The rainstorm duration (a) and frequency distribution (b) during D1, D2, and D3.

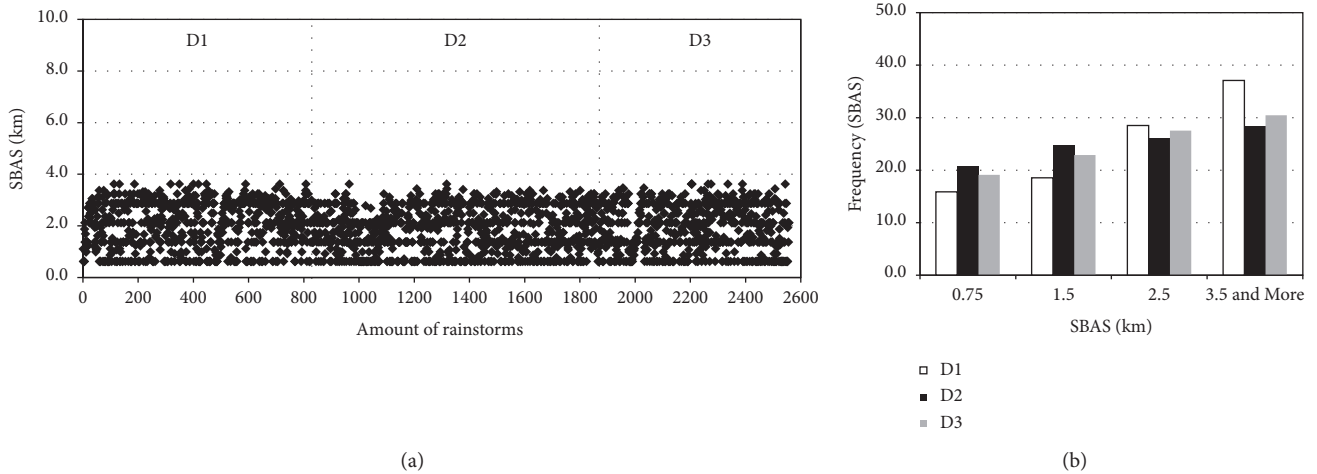


FIGURE 6: The rainstorm bases (a) and frequency distribution (b) during D1, D2, and D3.

Mean absolute error (MAE) is as follows:

$$MAE = \frac{1}{n} \sum_{i=1}^N |R_i - G_i|. \quad (4)$$

Bias is as follows:

$$B = \frac{\sum_{i=1}^N G_i}{\sum_{i=1}^N R_i}, \quad (5)$$

where G is 24-hour accumulations of rain gauges at station i in mm, R is 24-hour accumulations of radar rainfall computed using Z-R relationship at the point with coincided to rain gauge station i in mm, and N is the number of precipitation records.

Several Z-R relationships would be specified by the calculation of equations (2) to (5). Whichever relation provides the minimum of the four statistical measures will be selected as the most suitable relations for the study.

3. Result and Discussion

3.1. Rainstorm Properties of Tropical Storm PABUK from D1, D2, and D3

3.1.1. Rainstorm Duration (SDUR). SDUR is the time elapsed from the first radar reflectivity of 30 dBZ until the disappearance of precipitation. This study revealed that the average individual SDUR during D1 to D3 was around 0.9 to 1 hour, and almost all SDURs in D1 were less than 2 hours as in D2 and D3 tended to be longer duration than D1 as 1 to 3 hours. It is implied that the longer SDUR in D2 and D3 would extend the potential of rainfall intensity near the center of tropical storm PABUK in accordance with extremely heavy rainfall in D2 and D3 of rain gauge measurement when the rainstorm made landfall over southern Thailand. The time series and frequency distribution of SDUR on D1 to D3 are illustrated in Figure 5.

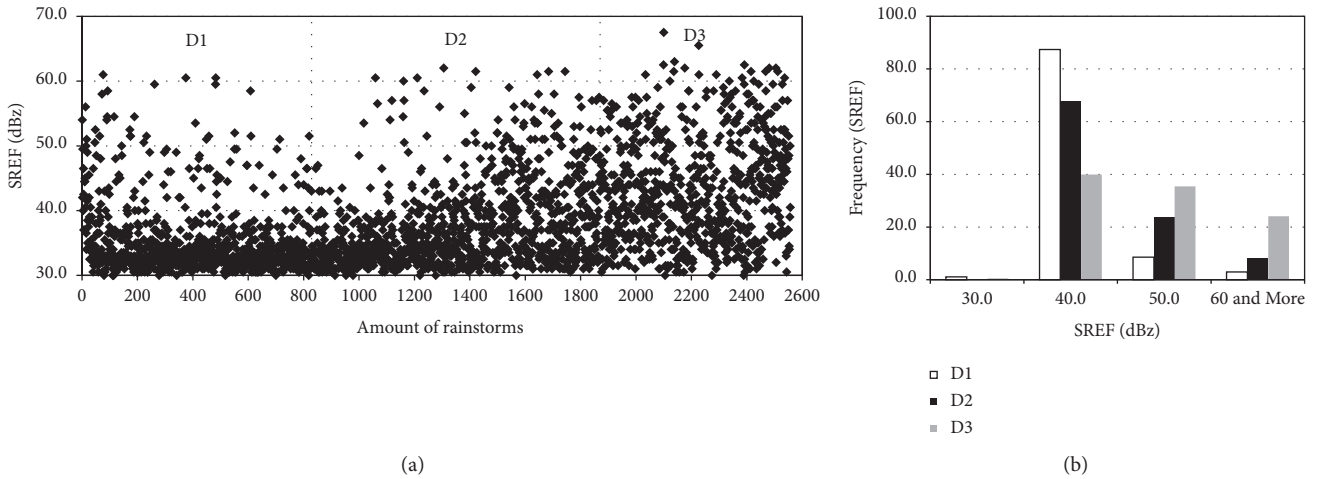


FIGURE 7: The rainstorm reflectivity (a) and frequency distribution (b) during D1, D2, and D3.

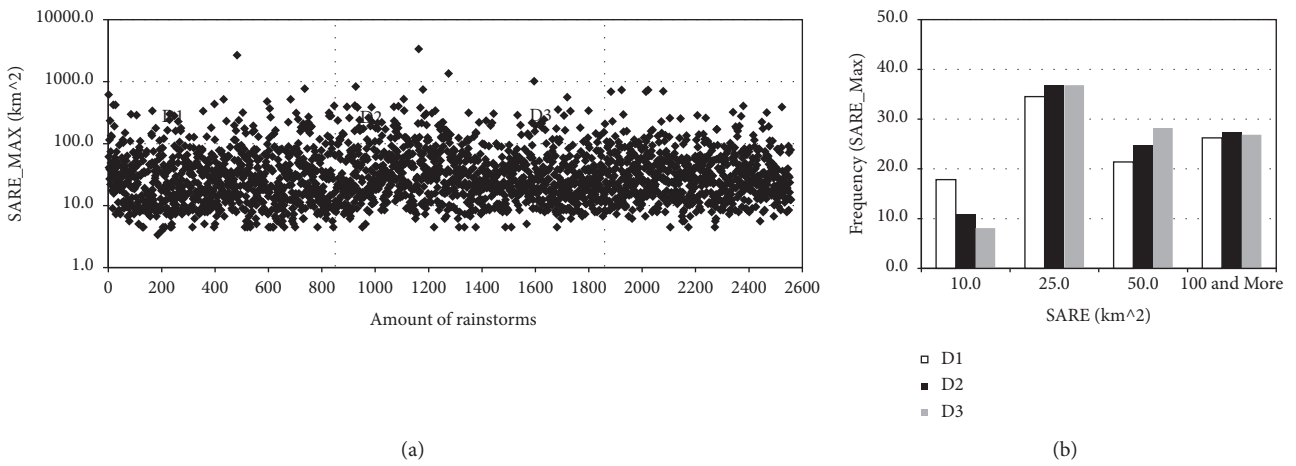


FIGURE 8: The rainstorm area (a) and frequency distribution (b) during D1, D2, and D3.

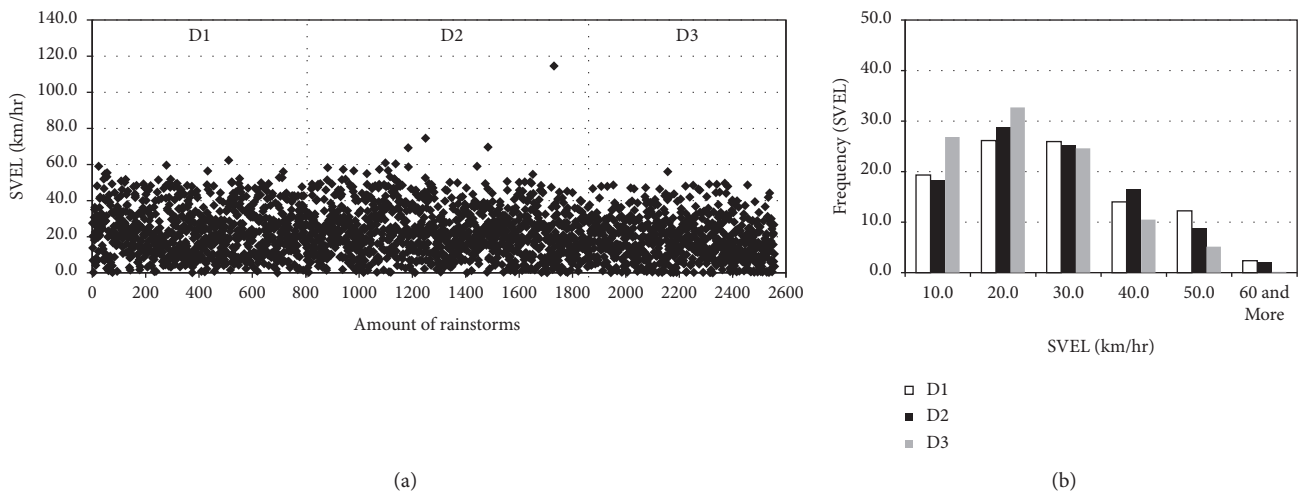


FIGURE 9: The rainstorm speed (a) and frequency distribution (b) during D1, D2, and D3.

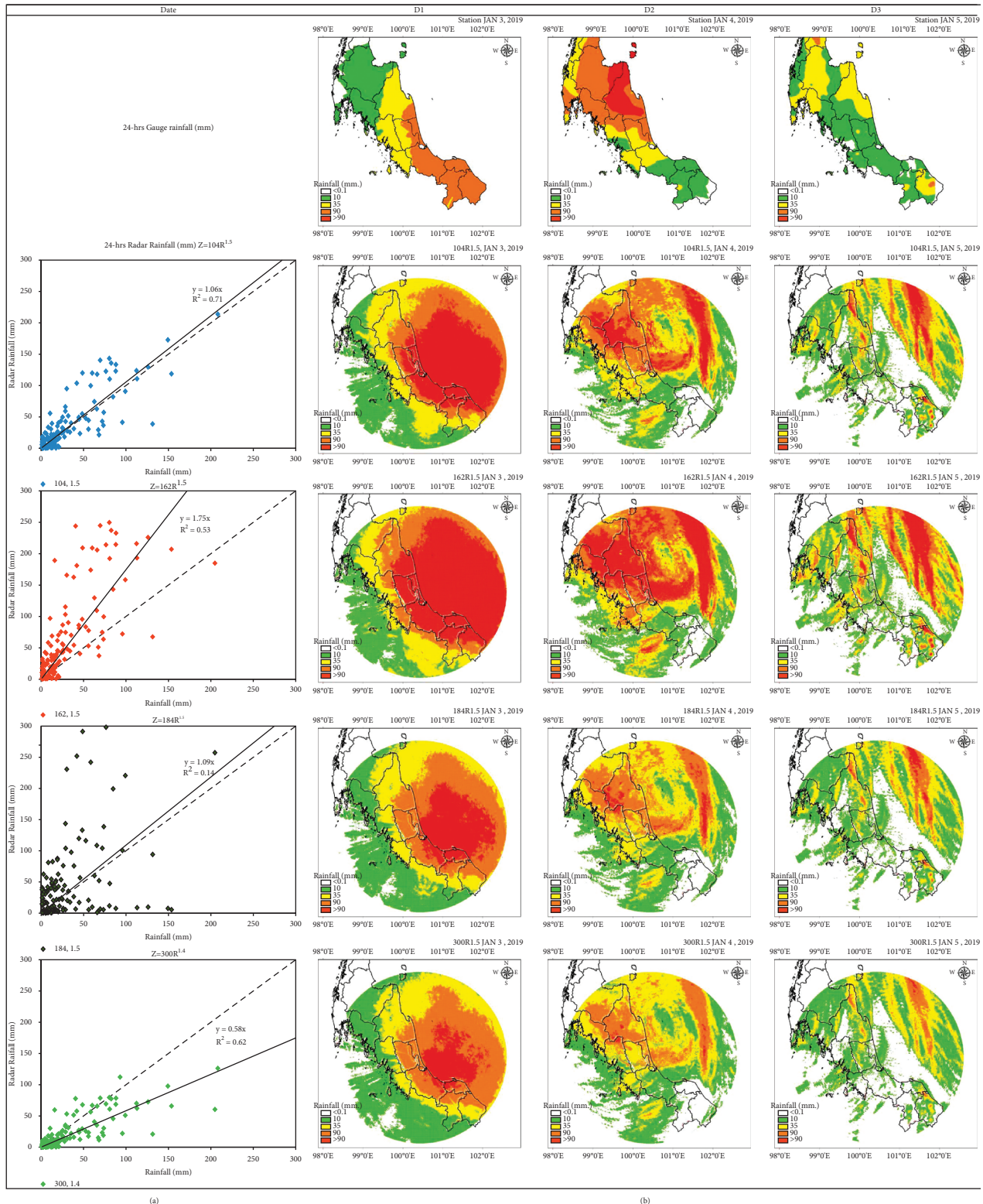


FIGURE 10: (a) Scatter plot between measured and estimated rainfall rate based on the different Z-R relationship of $Z = 184R^{1.5}$, $Z = 104R^{1.5}$, $Z = 162R^{1.5}$, and $Z = 300R^{1.4}$ and (b) time series plot of gauge rainfall and radar rainfall during D1, D2, and D3 by using the different Z-R relationship of $Z = 184R^{1.5}$, $Z = 104R^{1.5}$, $Z = 162R^{1.5}$, and $Z = 300R^{1.4}$.

3.1.2. *Rainstorm Bases (SBAS)*. The SBAS show the minimum height of radar reflectivity, as the minimum reflectivity threshold and altitude are determined as 30 dBZ and 0.6 km,

respectively. The results of this study showed that an average of SBAS is quite similar in D1, D2, and D3; between 1.8 and 2.0 km MSL, all events of occurred rainstorms are lower than

TABLE 3: Comparisons of the statistical measures gained from the different Z-R relationships of $Z = 184R^{1.5}$, $Z = 104R^{1.5}$, $Z = 162R^{1.5}$, and $Z = 300R^{1.4}$.

Z-R relationships	ME (mm)	RMSE (mm)	MAE (mm)	BIAS (G/R)
$Z = 104R^{1.5}$ (NEM)	-4.00	21.27	12.46	0.87
$Z = 162R^{1.5}$ (BULK)	-26.33	53.40	30.10	0.48
$Z = 184R^{1.5}$ (SWM)	-12.06	54.77	30.48	0.67
$Z = 300R^{1.4}$ (WSR)	9.34	24.38	12.94	1.52

4 km during PABUK occurrence. The time series and frequency distribution of SBAS on D1 to D3 are illustrated in Figure 6.

3.1.3. Rainstorm Reflectivity (SREF). The result from this study shows that on average maximum value, storm's reflectivity peaks that occurred during PABUK are distinguished among three days and tend to be higher SREF from D1, D2, and D3 as 35.0 dBZ, 38.6 dBZ, and 43.5 dBZ, respectively. Furthermore, the maximum SREF was found as a stronger reflectivity more than 60 dBZ which corresponds to the precipitation intensity and the development of PABUK tropical cyclone during D1 to D3 as well. The variation of SREF along three days in PABUK period is presented in Figure 7.

3.1.4. Rainstorm Area (SARE). The average SARE during PABUK period shows difference among D1, D2, and D3. The largest SARE from individual rainstorm is shown in D2 that is approximately 55.8 km². It is seen that the area of individual rainstorms of D2 when PABUK made landfall over Nakhon Si Thammarat province was found to be more potential rainstorms than D1 and D3 according to the extremely heavy rainfall in D2 occurred near the landfall point and inland of southern part. The D2 rainstorm composes of several large cells as well as they can also lead to larger areas of precipitation. The variation of these properties along three days in PABUK period is presented in Figure 8.

3.1.5. Rainstorm Speed (SVEL). The TITAN algorithm can provide the information of storm tracking and its movement. The results obtained from this study show that the average SVEL of D1, D2, and D3 was 23.1 kmhr⁻¹, 22.5 kmhr⁻¹, and 18.2 kmhr⁻¹, respectively. Most of SVEL has tend to be at lower speed when landfalling and passing through the land of southern Thailand. Investigating the maximum speed in D1, D2, and D3 as illustrated in Table 2, it is found that maximum SVEL of individual rainstorms was observed in D2 consistently with the report of maximum sustained wind during PABUK occurred period as well. The time series and frequency distribution of SVEL on D1 to D3 are illustrated in Figure 9.

3.2. Radar Rainfall Estimation. The radar estimated rainfall during PABUK landfall period was analyzed by the step as explained in the previous section. The result of estimated rainfall accumulation using Z-R relationship in four trials of (1) $Z = 184R^{1.5}$, (2) $Z = 104R^{1.5}$, (3) $Z = 162R^{1.5}$, and (4)

$Z = 300R^{1.4}$ is compared. The comparison of the 24 hr accumulated radar rainfall and the 24 hr accumulated gauge rainfall using the four trials of Z-R relationship is presented in Figure 10, and the statistical measures comparing these two sets of data are also calculated and summarized in Table 3.

Figure 10 shows the images of estimated daily radar rainfall attained from four Z-R relationships in D1, D2, and D3 and the scatter plot of the 24 hr accumulation of estimated radar rainfall attained from the different Z-R relationships and 24 hr accumulated gauge rainfall during D1 to D3. The estimated daily radar rainfall using the four trails of relation in D1, D2, and D3 was plotted as shown in the left side of radar images. From the scatter plot, it can be noted that the estimated radar rainfall accumulation is mostly higher than accumulated rain gauges except for $Z = 300R^{1.4}$ and also shows that $Z = 104R^{1.5}$ can provide the closest compared with the scatter plot of the other relations.

An agreement between estimated radar and gauge rainfall was examined using the statistical measures resulting from the four trials of Z-R relationships. The results show that the $Z = 104R^{1.5}$ is acceptable for overall statistical measures, with minimum of the four statistical measures, RMSE, ME, MAE, and BIAS, between the estimated radar and calculated rain gauge rainfall for the data sets in D1, D2, and D3. The calibrated Z-R relationship of $Z = 104R^{1.5}$ is therefore appropriate to be used for an estimation of accumulated radar rainfall in the tropical storm of PABUK.

4. Conclusions

The study presented the physical properties of rainstorm and radar-based rainfall estimation during tropical storm PABUK moving into the lower Gulf of Thailand and making landfall over southern Thailand which affected the southern regions with widespread heavy rainfall and flash floods. Derived from the data set of radar reflectivity and rain gauges during three days of PABUK, all storm properties were analyzed with TITAN, and estimated radar rainfall specified the appropriated Z-R relationship by the selected statistical measures. The results are shown as follows:

- (1) Three days (D1, D2, and D3) during the tropical storm of PABUK in order to investigate rainstorm of southern Thailand obtained the 5 properties of rainstorms by using TITAN, which provided the important analysis tool to identify rainstorms and their movement in this study. It is revealed that rainstorms were found to be the most effective clouds over southern region. In accordance with the

characteristics of rainstorms, it can be seen that rainstorms in D2 and D3 when PABUK made landfall over Nakhon Si Thammarat province, southern Thailand, show a longer lifetime, higher reflectivity, and larger rain-cells as well as it was found efficient in terms of rainfall amount than in D1 consistently with the records of high precipitation depth in southern in that periods.

- (2) The appropriated Z-R relationship acceptable for estimated radar rainfall during the tropical storm of PABUK in southern Thailand is $Z = 104R^{1.5}$, which provided the minimum of the four statistical measures (RMSE, ME, MAE, and BIAS) so far as the southern basin is concerned. The results should be especially useful in urban design problems as well as in hydrologic design problems during unusual cases such as extremely heavy rainfall from tropical storm on the southern basin.
- (3) These results are provided to assess the planning of water resources on a probability in a particular region or a basin and made to provide improvement of hydrometeorological relations that are pertinent to hydrological applications in the southern region and also capable of adapting to the other parts of Thailand. Relationships presented are subject to modification as additional data are collected in the heavy windy rainstorms, and further research is conducted.

Data Availability

All data used in the study are included in the manuscript.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors gratefully acknowledge the Department of Royal Rainmaking and Agricultural Aviation (DRRAA), Bangkok, Thailand, TMD, and HII for supporting the radar data sets and rain gauges through this research. The authors also appreciate the Royal-Rainmaking Atmospheric Observation Group for much helpfulness and providing the radar data sets and weather observation data used in this study.

References

- [1] P. Chantraket, C. Detyothin, and A. Suknarin, "Radar reflectivity derived rain-storm characteristics over northern Thailand," *EnvironmentAsia*, vol. 6, pp. 24–33, 2013.
- [2] P. Chantraket, S. Kirtsang, C. Detyothin, A. Nakburee, and K. Mongkala, "Characteristics of hailstorm over northern Thailand during summer season," *EnvironmentAsia*, vol. 8, pp. 101–114, 2015.
- [3] P. Chantraket, P. Intarachareon, and S. Kirtsang, "Analysis of rainstorms characteristics in eastern regions of Thailand," *International Journal of Applied Sciences and Innovations*, vol. 2016, no. 1, pp. 58–75, 2016.
- [4] H. Sauvgeot, "Rainfall measurement by radar: a review," *Atmospheric Research*, vol. 35, pp. 27–54, 1994.
- [5] J. Joss and A. Waldvogel, *Precipitation Measurement and Hydrology*, American Meteorological Society, Boston, MA, USA, 1990.
- [6] P. Smith, *Precipitation Measurement and Hydrology: Panel Report*, A. D. Meteorology, Ed., American Meteorological Society, Boston, MA, USA, 1990.
- [7] L. Peng, S. Xingyuan, and T. Wenwen, "A study on quantitative radar rainfall measurements by the method of set-air analysis," *Global Congress on Intelligent Systems*, 2009.
- [8] Y. Ma and V. Chandrasekar, "A hierarchical bayesian approach for bias correction of NEXRAD dual-polarization rainfall estimates: case study on hurricane irma in Florida," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, 2020.
- [9] K. Aydin and V. Giridhar, "C-band dual-polarization radar observables in rain," *Journal of Atmospheric and Oceanic Technology*, vol. 9, no. 4, pp. 383–390, 1992.
- [10] H. Chen, R. Cifelli, V. Chandrasekar, and Y. Ma, "A flexible bayesian approach to bias correction of radar-derived precipitation estimates over complex terrain: model design and initial verification," *Journal of Hydrometeorology*, vol. 20, no. 12, pp. 2367–2382, 2019.
- [11] Y. Gou, Y. Ma, H. Chen, and J. Yin, "Utilization of a C-band polarimetric radar for severe rainfall event analysis in complex terrain over eastern China," *Remote Sensing*, vol. 11, no. 1, p. 22, 2019.
- [12] W. Woodley and A. Herndon, "A raingage evaluation of the miami reflectivity-rainfall rate relation," *Journal of Applied Meteorology*, vol. 9, no. 2, pp. 258–264, 1970.
- [13] R. Suzana and T. Wardah, "Radar hydrology: new Z/R relationships for Klang river basin, Malaysia," vol. 8, pp. 248–251, in *Proceedings of the Radar Hydrology: New Z 2011 International Conference on Environment Science and Engineering IPCBEE*, vol. 8, pp. 248–251, IACSIT Press, Singapore, March 2011.
- [14] A. M. Hashem Albar and A. K. AL-Khalaf, "Radar rainfall estimation of a severe thunderstorm over jeddah," *Atmospheric and Climate Sciences*, vol. 5, pp. 302–316, 2015.
- [15] Y. Ma, Y. Zhang, D. Yang, and S. B. Farhan, "Precipitation bias variability versus various gauges under different climatic conditions over the Third Pole Environment (TPE) region," *International Journal of Climatology*, vol. 35, no. 7, pp. 1201–1211, 2015.
- [16] S. Kirtsang and P. Chantraket, "Investigation of Z-R relationships for monsoon seasons over southern Thailand," *Applied Mechanics and Materials*, vol. 855, pp. 159–164, 2017.
- [17] S. L. Barnes, "Report on a meeting to establish a common doppler radar data exchange format," *Bulletin American Meteorological Society*, vol. 61, pp. 1401–1404, 1980.
- [18] M. Dixon and G. Wiener, "TITAN: thunderstorm identification, tracking, analysis, and nowcasting—a radar-based methodology," *Journal of Atmospheric and Oceanic Technology*, vol. 10, no. 6, pp. 785–797, 1993.
- [19] J. T. Johnson, P. L. MacKeen, A. Witt et al., "The storm cell identification and tracking algorithm: an enhanced WSR-88D

- algorithm,” *Weather and Forecasting*, vol. 13, no. 2, pp. 263–276, 1998.
- [20] R. J. Potts, T. D. Keenan, and P. T. May, “Radar characteristics of storms in the sydney area,” *Monthly Weather Review*, vol. 128, no. 9, pp. 3308–3319, 2000.
- [21] R. A. Fulton, J. P. Breidenbach, D. J. Seo, D. A. Miller, and T. O’Brannon, “The WSD-88D rainfall algorithm,” *Weather Forecasting*, vol. 13, pp. 377–395, 1988.
- [22] C. W. Ulbrich and L. G. Lee, “Rainfall measurement error by WSR-88D radars due to variations in Z-R law parameters and radar constant,” *Journal of Atmospheric and Oceanic Technology*, vol. 16, pp. 258–264, 1999.
- [23] A. Seed, L. Sirivardena, X. Sun, P. Jordan, and J. Elliot, “On the calibration of Australian weather radars,” Technical Report 02/7, Cooperative Research Centre for Catchment Hydrology, Melbourne, Australia, 2002.

Research Article

Application of HMC-SS Method in Pipeline Reliability Analysis and Residual Life Assessment

Xin Lin ^{1,2} and Guojian Shao ¹

¹College of Mechanics and Materials, Hohai University, Nanjing 211100, China

²School of Architecture and Engineering, Anhui Polytechnic University, Wuhu 241000, China

Correspondence should be addressed to Guojian Shao; gjsao@hhu.edu.cn

Received 29 June 2021; Accepted 10 September 2021; Published 5 November 2021

Academic Editor: Jie Chen

Copyright © 2021 Xin Lin and Guojian Shao. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In this paper, the reliability analysis and residual life assessment model of gas pipelines with multiple corrosion pits are established. Aiming at the simulation evaluation of small failure probability of gas pipelines, a new method for reliability analysis and residual life assessment of gas pipelines with multiple internal corrosion pits is proposed, which is called the Hamiltonian Monte Carlo subset simulation (HMC-SS) method. Compared with the traditional MCS (Monte Carlo simulation) algorithm, the HMC-SS method has the advantages of less sampling, low cost, and high accuracy. And compared with the random walk SS method, the HMC-SS method can analyze the state space more efficiently and achieve faster convergence. In this paper, the HMC-SS method is applied to the reliability analysis and residual life assessment of gas pipeline engineering, and the sensitivity analysis of the random parameters affecting the failure probability of the pipeline is carried out. The results show that the corrosion rate, the depth of corrosion defects, and the wall thickness of the pipeline have great influence on the residual life of the pipeline, while the yield strength, working pressure, and the length of corrosion pits have no obvious influence on the failure probability and residual life of the pipeline. The analysis shows that the proposed HMC-SS method can be used as a reasonable tool for failure assessment of natural gas pipelines affected by corrosion to determine the remaining life of the pipeline system. This method provides a reliable theoretical basis for the integrity management of the gas pipeline.

1. Introduction

Pipeline transportation is an important part of gas production process. With the increasing service time of existing pipelines, the gas pipeline will inevitably be affected by an internal transport medium, external soil medium, and stray current in the long-term operation process, which will cause pipeline corrosion [1]. Corrosion will cause the pipeline wall thickness reduction, perforation, leakage, strength reduction, and cracking; serious cases will cause pipeline leakage or explosion, not only will cause serious casualties and major economic losses but also pollute the environment and cause bad social impacts. Therefore, once the oil and gas pipeline is put into operation for a period of time, the reliability analysis and remaining service life assessment of the pipeline are very important. At present, the assessment of corrosion pipelines

mainly includes ASME-B31 G [2], DNV RPF101 [3], API579 [4], and other standards. Many scholars have carried out reliability analysis and residual life assessment of corrosion pipelines through these standards. The most commonly used numerical simulation methods are a second-order moment method and MCS method. Teixeira et al. [5] used the first-order reliability method (FORM) and MCS simulation method to analyze the failure probability of the pipeline affected by internal pressure corrosion; Carr [6] points out that the failure probability obtained by MCS is more accurate and applicable than that obtained by FOSM because MCS is independent of the dimension of random variables and the complexity of limit state function; Li et al. [7] use MCS to evaluate the probability of pipeline instability. Although the first-order reliability method is simple to calculate, the error of the structural function with high

nonlinearity increases sharply [8]. The MCS method is widely used in the failure probability analysis of structures and residual life assessment due to its robustness and unbiased calculation results [9]. It can solve any model, especially when the sample size tends to be infinite, the accurate solution can be obtained, and the calculation results are often used as the standard to test the accuracy of other calculation methods. However, the sample size required by MCS is inversely proportional to the failure probability. This means that in order to ensure the accuracy and comprehensiveness of failure probability calculation, a large number of samples are needed when we use MCS to deal with the problem of high-dimensional small failure probability (for example, $\leq 10^{-3}$), and MCS is difficult to accept by the project. The failure analysis and residual life assessment of corrosion pipeline are high-dimensional and small failure probability problems. In order to solve this problem, an accurate and effective method is urgently needed. The improved MCS method, such as the SS simulation method, has obvious advantages in solving high-dimensional and small failure probability problems. The efficiency and robustness of the method are highly praised by many scholars [10–12]. Many researchers have applied the SS method to reliability analysis of bridges, buildings, and other engineering structures [13–16]. In the current practice of subset simulation, the most widely used method is to generate the required condition samples by various MCMC (Markov chain Monte Carlo) algorithms based on random walk. Au and Beck apply this method to the reliability problem of high-dimensional small failure probability, thus improving the calculation accuracy of subset simulation [17]. Miao and Ghosn use the subset simulation method based on MCMC to the safety and reliability analysis of structural systems [18]. Papaioannou et al. believe that the subset simulation method based on MCMC can effectively solve the problem of high-dimensional structural reliability [19]. Wang applies the subset simulation method based on MCMC to the reliability evaluation of steel bridges. The results show that this method is a good choice to solve the fatigue problems of nonlinear and multidimensional LSF [20]. Although this method has many advantages and is widely used, MCMC sampling method based on random walking makes Markov chain converge to the fixed distribution function $p(x)$, and the resulting conditional samples have high autocorrelation, so the accuracy of simulation results is poor and the efficiency is still very low [21]. In order to overcome these shortcomings and solve the practical problems of engineering, a new MCMC algorithm, which is more efficient and accurate, is adopted in this paper. HMC is used to calculate the future state of Markov chain rather than probability distribution by using the concept of dynamics in the physical system [22]. In this way, the state space can be analyzed more efficiently and the convergence can be achieved faster.

In recent years, HMC has been widely used and developed rapidly and has made remarkable achievements in various statistical applications [23–25]. The HMC method has been applied to Bayesian analysis and reliability analysis of structural engineering problems perfectly [26–28]. There is no research on this aspect in the reliability analysis and residual life assessment

of corrosion pipelines. In this paper, the system reliability analysis method is used to evaluate the pipelines with multiple corrosion pits. The model of reliability analysis and residual life assessment of corrosion pipelines based on HMC-SS method is established. The results based on the HMC-SS method are compared with those of SS and MCS methods.

The structure of this paper is as follows. In Section 2, the reliability analysis model of the gas pipeline based on reliability theory is established. In Section 3, the SS method based on random walking is reviewed. The fourth section describes the calculation details and flow of HMC-SS algorithm in detail. Section 5 proves the effectiveness of the method by a specific engineering example and further analyzes the sensitivity of the random variables that affect the remaining life of the gas pipeline. The conclusion and prospect are given in Section 6.

2. Reliability Analysis Model of Gas Pipelines Based on Reliability Theory

Based on the reliability theory, the state function of pipeline failure due to corrosion can be expressed by the following formula:

$$G = F_f - F_0, \quad (1)$$

where F_f is the failure pressure of the corroded pipeline and F_0 is the actual operating pressure of the pipeline. The failure probability of corroded pipeline can be expressed by the following formula:

$$P_f = P[G = F_f - F_0 \leq 0], \quad (2)$$

$$\Phi(\beta) = 1 - P_f, \quad (3)$$

where P is the probability of the event, Φ is the cumulative distribution function, and β is the reliability index of the structure.

For gas pipelines with corrosion defects, according to the revised ASME-B31 G standard measurement, the failure stress formula of corrosion pipeline is as follows:

$$\sigma_p = \sigma_y \frac{1 - (A/A_0)}{1 - (A/A_0)F}, \quad (4)$$

where σ_y is the flow stress of the pipeline material, A is the projected area of the corrosion defect on the side, A_0 is the original cross-sectional area of the pipeline at the corrosion defect, F is the Folias expansion factor, which is related to the corrosion pipe outer diameter D , wall thickness t , and corrosion defect length L , namely:

$$F = \sqrt{1 + 0.6275 \left(\frac{L^2}{Dt} \right) - 0.003375 \left(\frac{L^4}{D^2 t^2} \right)} \quad \frac{L^2}{Dt} \leq 50, \quad (5)$$

$$F = 0.032 \left(\frac{L^2}{Dt} \right) + 3.3 \quad \frac{L^2}{Dt} > 50.$$

In equation (4), $\sigma_y = k\sigma_s$ (σ_s is the yield stress of pipeline material), where k is the coefficient. For steel tubes, k is

usually 1.1 or 1.15, and in this paper, k is 1.15. $A_0 = Lt$ and $A = 0.85dL$ (d is the maximum depth of the corrosion pit and L is the maximum length of the corrosion pit). To predict the strength of a pipe at time T , it is necessary to estimate the corrosion growth rate. In [29, 30], a reasonable linear model is proposed to predict the corrosion growth of steel pipes and estimate the size of corrosion pits at time T .

$$\begin{cases} d = d_0 + \nu_d(T - T_0), \\ L = L_0 + \nu_L(T - T_0). \end{cases} \quad (6)$$

In equation (6), d_0, L_0 denotes the depth and length of corrosion defects detected in T_0 , respectively. d and L denote the depth and length of corrosion defects after service T , respectively; ν_d denotes the radial corrosion rate; and ν_L denotes the axial corrosion rate. Substituting (6) into (4), the failure pressure F_f of the pipeline with corrosion defects can be obtained as follows:

$$F_f = \frac{2t\sigma_p}{D} = \frac{2.3t\sigma_s}{D} \times \frac{1 - [d_0 + \nu_d(T - T_0)]/t}{1 - [d_0 + \nu_d(T - T_0)]/Lt}. \quad (7)$$

Since the failure pressure of the corroded pipeline is relatively independent of the actual operating pressure (F_0) of the pipeline, the limit state function G of the corroded pipeline is established based on the reliability theory as

$$G = F_f - F_0 = \frac{2.3t\sigma_s}{D} \times \frac{1 - [d_0 + \nu_d(T - T_0)]/t}{1 - [d_0 + \nu_d(T - T_0)]/Lt} - F_0. \quad (8)$$

When $G > 0$, the structure is in a reliable state; $G < 0$ indicates that the structure is in a failure state; $G = 0$ means that the structure is in the limit state.

3. SS Algorithm Based on Random Walk MCMC Sampling

The basic idea of SS algorithm based on random walk MCMC sampling is to convert the small failure probability into the product of a series of large conditional failure probability events by introducing reasonable intermediate failure events. In the failure region for the functional function $G(X)$: $E = \{G(X) < b\}$, a series of thresholds can be introduced as $b_1 > b_2 > \dots > b_m = b$. Failure events with nested relationships are composed of these thresholds. $E_k = \{G(X) < b_k, k = 1, 2, \dots, m\}$, and m is the total number of intermediate events. At this time, $E_1 \subset E_2 \subset \dots \subset E_m = E$. According to conditional probability theory, the target failure probability P_f can be obtained as

$$P_f = P(E_m) = P(E_1) \prod_{k=2}^m P\left(\frac{E_k}{E_{k-1}}\right). \quad (9)$$

To facilitate calculation, the intermediate conditional probability is set to a constant value P_0 and then the threshold value b_m can be determined by the adaptive method. The traditional subset simulation method has a detailed process description in [31, 32]. The approximate

value of the final failure probability can be obtained by the following equation:

$$\widehat{P}_f = \frac{P_0^{m-1}}{N} N_{E_m} \approx P_f, \quad (10)$$

where N_{E_m} is the number of samples falling into the last layer.

4. Reliability Analysis Model of Gas Pipelines Based on HMC-SS Algorithm

HMC-SS is an improvement on the traditional SS algorithm. This method combines Hamiltonian dynamics with MCMC algorithm and applies it to subset simulation, which has the advantages of faster convergence, higher accuracy, and better efficiency.

4.1. Principle of Hamiltonian Dynamics. Hamiltonian dynamics describes the time evolution of the system according to the position vector q and momentum vector p . The total energy of the system H is a function of q and p , namely:

$$H(q, p) = U(q) + K(p), \quad (11)$$

where $U(q)$ is the potential energy and $K(p)$ is the kinetic energy.

Hamiltonian dynamics has several important properties: (1) reversibility, (2) conservation of the Hamiltonian, (3) volume preservation, and (4) symplecticness. These properties are the most important conditions for Hamiltonian dynamics to be applied to Markov chain Monte Carlo updating, and the analysis and proof process are described in detail in references [28, 33]. In order to connect the Hamiltonian dynamics with the Markov chain Monte Carlo method, firstly, the random variables x are regarded as q (i.e., $x \equiv q$) of the Hamiltonian system and p as independent variables with the same dimension, so x and p together form the extended initial position space; then, the position momentum state space of the Hamiltonian system is obtained. The potential energy function $U(x)$ can be determined by the target probability density function $\pi(x)$:

$$U(x) = -\log\pi(x). \quad (12)$$

In general, the kinetic energy function $K(p)$ can be defined as follows:

$$K(p) = \frac{p^T M^{-1} p}{2}, \quad (13)$$

where M is a positive definite, symmetric "mass" matrix, generally taking the scalar product of the unit matrix (the scalar matrix). Equation (11) can be rewritten as follows:

$$H(x, p) = U(x) + K(p). \quad (14)$$

We use the knowledge of statistical mechanics to connect Hamiltonian functions $H(x, p)$ and $\pi(x)$. According to the energy function $E(\theta)$, we can define a regular distribution:

$$\pi(\theta) = \frac{1}{Z} e^{-E(\theta)}, \quad (15)$$

where Z is the regularization coefficient and can guarantee $\int \pi(\theta) d\theta = 1$. Then, $E(\theta)$ is

$$E(\theta) = H(x, p) = U(x) + K(p). \quad (16)$$

Then, the regular distribution of Hamiltonian energy function can be expressed as

$$P(x, p) \propto e^{-H(x,p)} \propto e^{-U(x)} e^{-K(p)} \propto P(x)P(p). \quad (17)$$

It can be seen from equation (14) that $P(x, p)$ can be decomposed into the product of $P(x)$ and $P(p)$. It can be seen that these two variables are independent of each other. So, we can sample from the joint probability density function of x and p . In the Hamiltonian system, $H(x, p)$ is a constant value, which describes how the kinetic energy and potential energy are transformed into each other in the process of system motion. It can be analyzed quantitatively in the form of differential equation as follows:

$$\begin{cases} \frac{dx_i}{dt} = \frac{\partial H}{\partial p_i}, \\ \frac{dp_i}{dt} = -\frac{\partial H}{\partial x_i}. \end{cases} \quad (18)$$

The rationality and validity of applying Hamiltonian equation to the MCMC method are explained in [32, 34]. The common methods of discrete Hamiltonian equation are Euler method, improved Euler method, jump point method, and so on [34]. These methods have been described in detail in reference [34]. Because the jump point method is simple, reversible, and retains the symplectic structure of the phase space, it has high efficiency and high accuracy in solving Hamiltonian equation. In this paper, the jump point method is used, and the formula is as follows:

$$\begin{cases} p_i\left(t + \frac{\varepsilon}{2}\right) = p_i(t) - \left(\frac{\varepsilon}{2}\right) \frac{\partial U}{\partial x_i}(x(t)), \\ x_i(t + \varepsilon) = x_i(t) + \varepsilon \frac{p_i(t + \varepsilon/2)}{m_i}, \\ p_i(t + \varepsilon) = p_i\left(t + \frac{\varepsilon}{2}\right) - \left(\frac{\varepsilon}{2}\right) \frac{\partial U}{\partial x_i}(x(t + \varepsilon)), \end{cases} \quad (19)$$

where ε is the time step, $L = (\tau/\varepsilon)$, L is the number of iterations, and τ is the length of the path. The efficiency of the jump point method is very dependent on the selection of ε and L . In this paper, when the average acceptance rate is about 65%, the corresponding L and ε are selected. Generally, it is assumed that the value between 60% and 80% is the best, which is determined by dual algorithm. It is described in [34–36] in detail. Hamiltonian function is applied to the MCMC method, which is called HMC algorithm in this paper. HMC algorithm is used to extract samples from conditional probability density function $f(x/E_{k-1})$. The algorithm flow is as follows:

- (1) An initial momentum P_{init} (where M is the unit matrix) is generated from the normal distribution $N(0, M)$.
- (2) The initial momentum P_{init} and the position x_{init} of the seed sample are used as the initial conditions to generate a new state (x^*, p^*) according to equation (16).
- (3) Receiving calibration: if $x^* \in E_k$ and $\text{rand} < \min[1, \exp(-H(x^*, p^*)) + H(x_{\text{init}}, P_{\text{init}})]$, where $\text{rand} \sim U([0, 1])$, receive the suggested sample x^* as the next state; otherwise x^* cannot be the next state, and the next state is still the current state x_{init} , end.

HMC sampling can explore the state space more effectively than random walk MCMC sampling, which greatly improves the computational efficiency. The following is a binary Gaussian distribution, where the mean value of position variable is 0, and the covariance matrix is $[1, 0.9; 0.9, 1]$, the momentum variable is 0, and the covariance matrix is $[1, 0; 0, 1]$. At this time, Hamiltonian function can be defined as the following formula:

$$H(x, p) = x^T \sum \frac{-1}{2} x + p^T \frac{P}{2}, \quad (20)$$

$$\sum = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}.$$

From Figures 1 and 2, we can see that HMC algorithm has higher probability of accepting samples than traditional random walk sampling methods and can analyze state space more efficiently, thus achieving faster convergence.

4.2. The Flow of Hamiltonian Monte Carlo Subset Simulation Algorithm. The flow of Hamiltonian Monte Carlo subset simulation algorithm is as follows:

- (1) The total number of samples N and conditional failure probability P_0 are defined. Let $Ns = NP_0$ and $Nt = N - Ns$.
- (2) According to the probability density function of the model, the Monte Carlo simulation method is directly used to generate N independent samples.
- (3) Take the samples into the limit function to get the response value of the corresponding samples and sort them from small to large to get $\{G_j^{(1)}, j = 1, 2, \dots, N\}$ and the corresponding sample sequence $\{x_j^{(1)}, j = 1, 2, \dots, N\}$, respectively.
- (4) Take $b_1 = G_{Ns}$, then $P(E_1) = P_0$; let $k = 2$.
- (5) Take $\{x_j^{(k-1)}, j = 1, 2, \dots, Ns\}$ as the “seed” sample and generate Nt groups of new samples $\{x_l, l = 1, 2, \dots, Nt\}$ that meet the conditional probability density function $f(x/E_{k-1})$ according to the HMC algorithm. Specific steps can be carried out as follows:

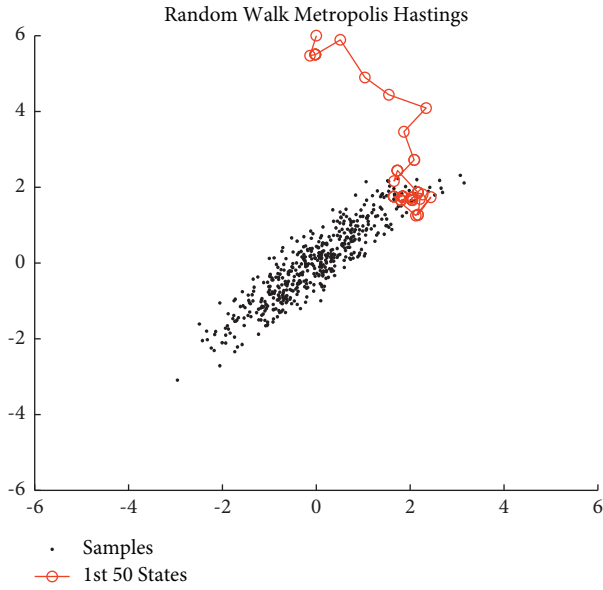


FIGURE 1: 200 iterations of random walk MH method (acceptance probability $\alpha = 0.87$).

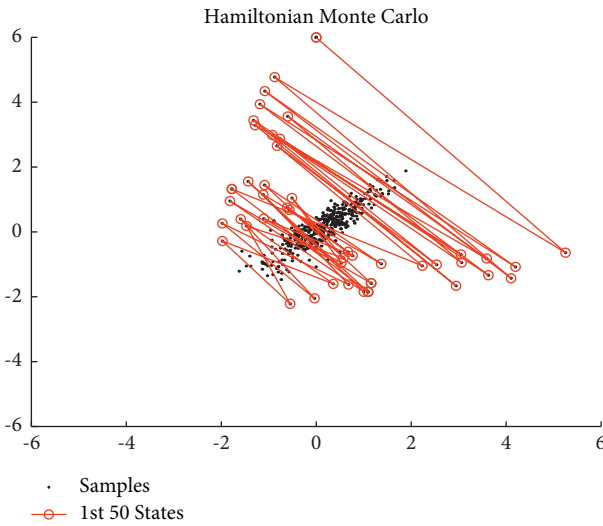


FIGURE 2: HMC method with 200 iterations ($L = 200, \epsilon = 0.3$) (acceptance probability $\alpha = 0.97$).

- (a) Through the “seed” samples, the Hamiltonian jump point method is used to generate Nt groups of new samples.
- (b) The original Ns groups of “seed” samples and Nt groups of new samples are brought into the limit function to obtain the corresponding response value.
- (c) Reorder the response values in ascending order to get $\{G_j^{(k)}, j = 1, 2, \dots, N\}$ and the corresponding sample sequence $\{x_j^{(k)}, j = 1, 2, \dots,$

$N\}$. Take $b_k = G_{Ns}^{(k)}$ and get the intermediate conditional probability $p_k = p_0$.

- (6) If at least $\{G_j^{(k)}\}$ response values in Ns reach the threshold, i.e., $G_{Ns}^{(k)} = b$, let $k = 2$ go directly to (7); otherwise, let $k = k + 1$ go to (5).
- (7) The number of samples falling into the failure domain E_m is calculated, i.e.,

$$P_f \approx \frac{P_0^{m-1}}{N} N_{E_m}. \quad (21)$$

When a pipeline has multiple corrosion points, the failure occurring at each corrosion point constitutes the total failure of the pipeline. Therefore, the series system is more suitable for the failure assessment of corroded pipelines. The failure probability $P_{f,s}$ of pipelines can be estimated by using the reliability method of a series system [37, 38]:

$$LB_{P_{f,s}} = \max[P_{f,i}] \leq P_{f,s} \leq 1 - \prod_{i=1}^n [P_{f,i}] = UB_{P_{f,s}}, \quad (22)$$

where $P_{f,i}$ represents the failure probability of the i th corrosion pit, which can be calculated by (21). n is the total corrosion points, and $LB_{P_{f,s}}$ and $UB_{P_{f,s}}$ represent the lower and upper ascertainment of the failure probability of the system, respectively. According to the requirements of the geographical location of the pipeline on the risk level, when the failure probability of the corroded pipeline is greater than the acceptable failure probability over time, it is considered the pipeline damage; that is, the remaining life of the corroded pipeline is calculated.

5. Numerical Examples

Taking the secondary high-pressure gas pipeline in a corrosion area as an example, under the assumption of independent random variables, the pipeline has been running for 20 years. The pipe specifications are as follows: 273×7 , 20# steel, yield strength of 245 MPa, and maximum allowable working pressure of 0.7 MPa. And, one of the sections is sampled. The results show that there are three corrosion pits, and the gas pipeline parameters and geometric features of the corrosion pits are shown in Tables 1 and 2.

According to the target reliability of the API579 standard (see Table 3), the geographical location of the pipeline is in the middle risk area and the allowable failure probability of the corroded pipeline is 10^{-3} . When the failure probability of the pipeline is greater than 10^{-3} , the initial time is the remaining life of the pipeline.

In the corrosion failure process of a gas pipeline, the failure probability of the corroded pipeline is calculated by the upper bound of equation (22). To improve the accuracy of the calculation, the sample number of the MCS method is 10^9 and its result is approximate to the analytical result.

TABLE 1: Pipeline parameters.

Random variable	Distribution type	μ	Std
Diameter, D (mm)	Normal distribution	273	5.46
Wall thickness, t (mm)	Normal distribution	7	0.14
Pipeline operating pressure, F_0 (MPa)	Normal distribution	0.7	0.07
Pipe yield strength, σ_s (MPa)	Normal distribution	245	17.15
Radial corrosion rate, ν_d ($\text{mm} \cdot \text{a}^{-1}$)	Normal distribution	0.229	0.0229
Axial corrosion rate, ν_l ($\text{mm} \cdot \text{a}^{-1}$)	Normal distribution	0.229	0.0229

TABLE 2: Geometry of pipeline corrosion pits.

	Corrosion pit 1		Corrosion pit 2		Corrosion pit 3	
	Length (mm)	Depth (mm)	Length (mm)	Depth (mm)	Length (mm)	Depth (mm)
Mean	115	4.64	150	4.0	200	3.6
Standard deviation	11.5	0.928	15	0.8	20	0.72

TABLE 3: Reliability of pipeline operation target.

Location category	Low risk area (category I areas)	Medium risk area (category II areas)	High risk area (category III and IV areas)
Acceptable failure probability	10^{-2}	10^{-3}	10^{-5}

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^m (Pf_{\text{MCS}} - Pf(i))^2}{m}} \quad (23)$$

Among them, Pf_{MCS} is the failure probability of MCS ($N = 10^6$), $Pf(i)$ is the failure probability calculated by simulation method, and m is the number of simulations.

In order to obtain clear observation results, Figure 4 takes logarithmic coordinates, and it can be seen from Figure 4 that RMSE of HMC-SS and MCS methods is smaller than that of SS and MCS methods. This shows that compared with the SS method, the HMC-SS method has higher accuracy and more effective calculation failure probability.

According to Figure 3, the pipeline runs for another 17 years, if the failure probability is 8.0×10^{-4} ; the pipeline continues to run 18 years, if the failure probability is 2.59×10^{-3} ; the pipeline residual life is 17 years because the failure probability of the pipeline is over the pipeline allowance of 10^{-3} . The simulation results of SS method and HMC-SS method are also 17 years. The results of the three methods are in good agreement.

It can be seen from Table 4 that the failure probability calculated by the MCS method, SS method, and HMC-SS method for another 18 years of pipeline operation is 5.29×10^{-3} , 1.4×10^{-3} , and 2.08×10^{-3} . The corresponding running time is 1.0561×10^2 seconds, 13.818 seconds, and 11.281 seconds.

Through the comparative analysis, it can be seen that the MCS calculation for the pipeline with three corrosion pits needs a lot of calculation cost, about 0.915×10^2 seconds. When using SS, HMC-SS algorithm only needs 1% of MC samples to achieve the same accuracy, and the running time is less than 8.49% and 8.77% of MCS algorithm. Therefore, the SS and HMC-SS algorithms can be more easily applied to the system reliability analysis and life prediction with multiple corrosion points.

RMSE (root mean square error) is used to measure the deviation between the simulated value and the accurate value. The smaller the RMSE is, the higher the accuracy of the simulation value is. In this paper, the root mean square error between the failure probability of 1000:1000:5000 (the number of samples is 1000 to 5000, with a step of 1000) and MCS ($N = 10^6$) is calculated by equation (23), respectively:

It can be seen from Figure 5 that with the increase of the working pressure of the gas pipeline, the failure probability of the pipeline increases and the remaining life decreases gradually. When the working pressure increases from 0.25 MPa to 1.25 MPa, the corresponding remaining life decreases from 22 years to 16 years. It can be seen from Figure 6 that the change of corrosion rate has greater influence on the remaining life of the pipeline, when the corrosion rate ν_d and ν_l increased from $0.113 \text{ mm} \cdot \text{a}^{-1}$ to $0.458 \text{ mm} \cdot \text{a}^{-1}$, the service life of gas pipeline is reduced from 23 years to 14 years. Without replacing the gas pipeline, the corrosion treatment should be strengthened actively to delay the corrosion rate of the pipeline and ensure the safety of the gas pipeline.

It can be seen from Figure 7 that the yield strength σ_s of the gas pipeline has little influence on the remaining life of the pipeline, while it can be seen from Figure 8 that the geometric shape of the gas pipeline has a significant influence on the remaining life of the pipeline. The remaining life of pipe decreases sharply with the decrease of pipe wall thickness, which is also consistent with the actual situation. The thicker the pipe wall is, the stronger the corrosion resistance is, and the less likely it is to fail.

To make a more detailed and intuitive analysis of the influence of random variables on the failure probability and the remaining service life of gas pipelines, the reliability sensitivity of the mean value of random variables is shown in Figure 9. The paper [38] provides the reliability sensitivity

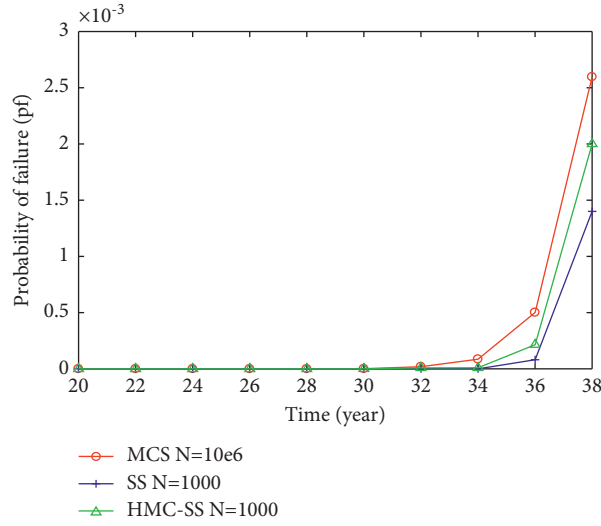


FIGURE 3: Curve of failure probability and service life of corroded pipeline.

TABLE 4: Comparison of results between HMC-SS, SS, and MC methods.

Simulation method	Sample number	Elapsed time (s)	Probability of failure (pf)	Running time (T) (years)
MCS	10^6	0.915×10^2	4.12×10^{-3}	18
SS ($p_0 = 0.1$)	1000	13.818	1.4×10^{-3}	18
HMC-SS	1000	11.281	2.08×10^{-3}	18

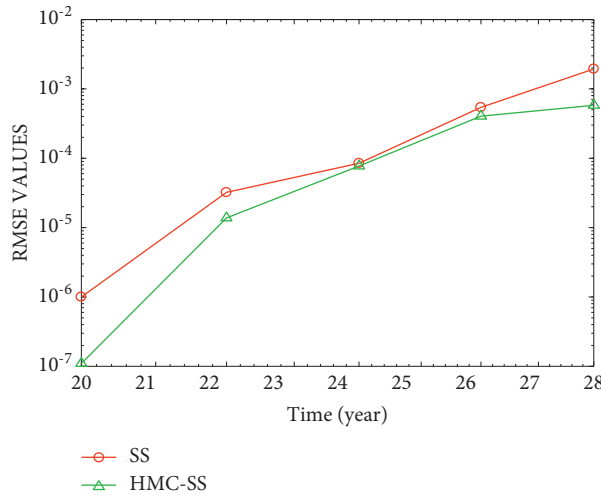


FIGURE 4: RMSE of SS method and HMC-SS method compared with MCS method, respectively.

computation formula as follows: $S_X = (\partial P_f / \partial X)$, from graphics can be very intuitive found that the most important factors affecting the failure probability and remaining service life of the gas pipeline are the corrosion rate, wall thickness,

corrosion pit depth, and gas transmission pressure. The yield strength of the pipe material and the length of corrosion pits have almost negligible effects on the failure probability and remaining service life.

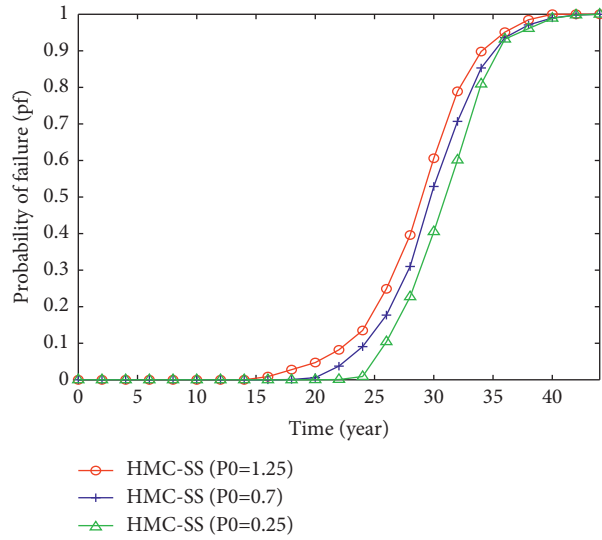


FIGURE 5: Failure probability at different working pressures.

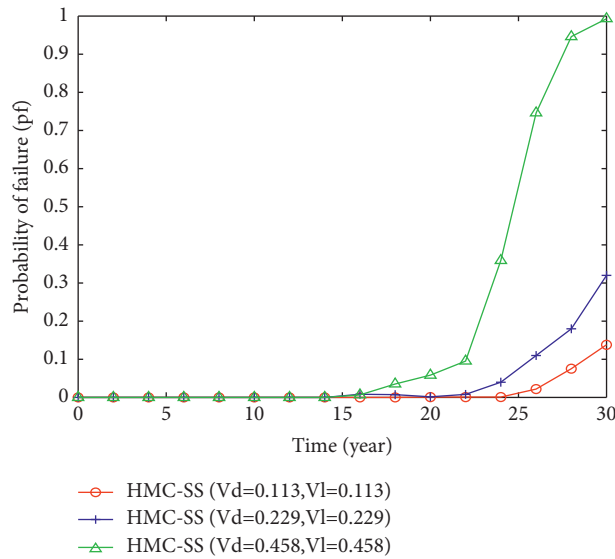


FIGURE 6: Failure probability at different corrosion rates.

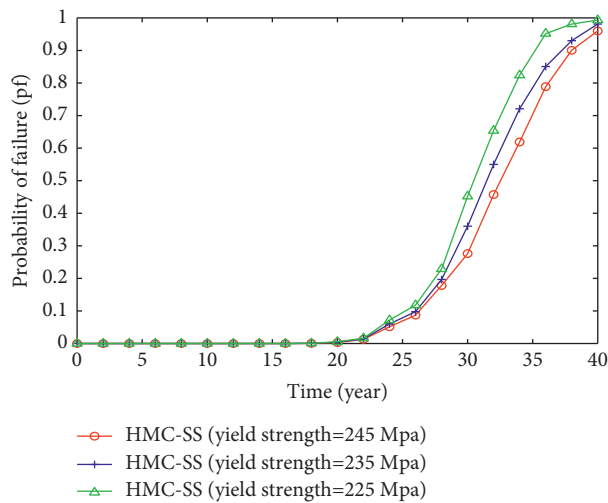


FIGURE 7: Failure probability at different working pressures.

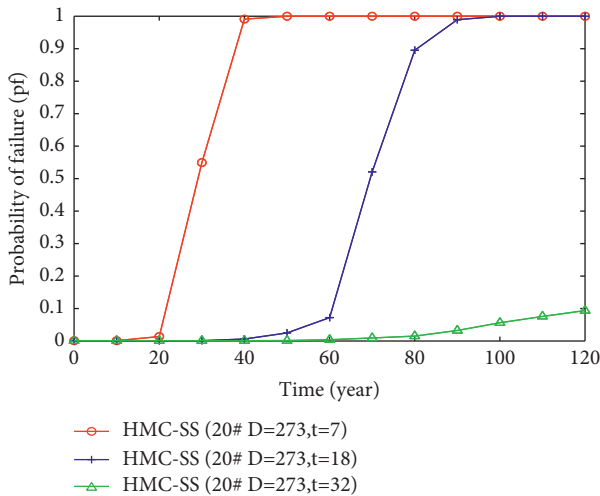


FIGURE 8: Failure probability at different corrosion rates.

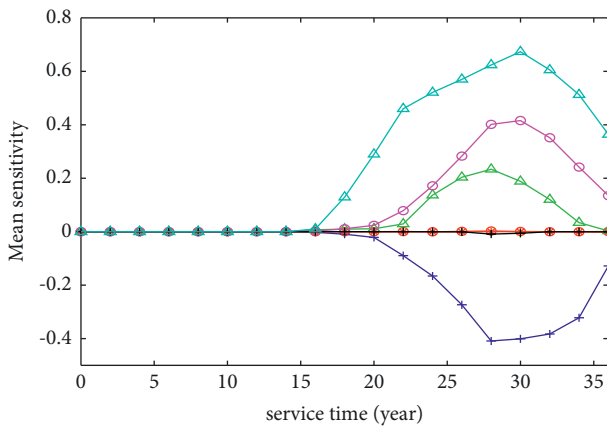


FIGURE 9: Average sensitivity curve.

6. Conclusions

In view of the characteristics of multimode failure and small failure probability in the residual life of urban gas transmission and distribution pipelines, a reliability residual life calculation model of gas pipelines with multiple corrosion pits is established in this paper. The MCS, SS, and HMC-SS methods are used to calculate the residual life of gas pipelines, and the influence of relevant parameters is discussed.

- (1) In the case of a small failure probability problem with multiple corrosion pits, HMC-SS has advantages of less sampling, saving time, and high calculation accuracy.
- (2) The analysis of sensitivity shows that the corrosion rate, wall thickness, corrosion defect depth, and gas transportation pressure of the pipeline have obvious influences on the failure probability and the remaining service life of the gas pipeline, while the yield strength of the pipe and the length of the corrosion pit have little influence on the failure probability and the remaining service life of the pipeline.

- (3) Considering that the corrosion rate and wall thickness of the pipeline have an obvious influence on the remaining service life of the gas pipeline, it is suggested to select the appropriate pipe type according to the geographical location of the pipeline to meet the safety requirements and make it economical.
- (4) According to the sensitivity of factors affecting the remaining life of gas pipelines, it is necessary to prioritize the maintenance or replacement of in-service pipelines and select the materials for laying pipelines to provide theoretical guidance.

Data Availability

Some or all data, models, or codes generated or used during the study are available from the corresponding author upon request.

Conflicts of Interest

The authors have no conflicts of interest to declare.

Acknowledgments

The work described in this paper was supported by the Anhui Provincial Natural Science Foundation (No. 2008085QA22).



References

- [1] M. K. Totlani and S. N. Athavale, "Electroless nickel for corrosion control in chemical, oil and gas industries: corrosion reviews," *Corrosion Reviews*, vol. 18, no. 2/3, pp. 155–180, 2000.
- [2] American Society of Mechanical Engineers, *ASME B31G-2009, Manual for Determining the Remaining Strength of Corroded Pipeline*, ASME B31 Committee, New York, NY, USA, 2009.
- [3] Det Norske Veritas, *DNV -RP -F101 1999, Recommended Practice for Corroded Pipelines*, Det Norske Veritas, Oslo, Norway, 1999.
- [4] API 579, *2000 Recommended Practice for Fitness for Service*, American Petroleum Institute, no. 6, Washington, DC, USA, 2000.
- [5] A. P. Teixeira, C. Guedes Soares, T. A. Netto, and S. F. Estefen, "Reliability of pipelines with corrosion defects," *International Journal of Pressure Vessels and Piping*, vol. 85, no. 4, pp. 228–237, 2008.
- [6] P. Carr, "Riser and pipeline corrosion risk assessment," in *Proceedings of the Offshore Technology Conference-Asia. Offshore Technology Conference*, Houston, Texas, USA, May 2014.
- [7] X. Li, G. Chen, H. Zhu, and R. Zhang, "Quantitative risk assessment of submarine pipeline instability," *Journal of Loss Prevention in the Process Industries*, vol. 45, pp. 108–115, 2017.
- [8] G. I. Schuëller and R. Stix, "A critical appraisal of methods to determine failure probabilities," *Structural Safety*, vol. 4, no. 4, pp. 293–309, 1987.
- [9] R. Y. Rubinstein and D. P. Kroese, *Simulation and the Monte-Carlo Method*, Wiley, New York, NY, USA, 2007.

- [10] L. R. Khan and K. F. Tee, "Risk-cost optimization of buried pipelines using subset simulation," *Journal of Infrastructure Systems*, vol. 22, no. 2, Article ID 04016001, 2016.
- [11] Y. Li, Y. Zhang, and D. Kennedy, "Reliability analysis of subsea pipelines under spatially varying ground motions by using subset simulation," *Reliability Engineering & System Safety*, vol. 172, pp. 74–83, 2018.
- [12] K. F. Tee, L. R. Khan, and H. Li, "Application of subset simulation in reliability estimation of underground pipelines," *Reliability Engineering & System Safety*, vol. 130, pp. 125–131, 2014.
- [13] S. K. Au and J. L. Beck, "Subset simulation and its application to seismic risk based on dynamic analysis," *Journal of Engineering Mechanics*, vol. 129, no. 8, pp. 901–917, 2003.
- [14] J. Ching, J. L. Beck, and S. K. Au, "Hybrid Subset Simulation method for reliability estimation of dynamical systems subject to stochastic excitation," *Probabilistic Engineering Mechanics*, vol. 20, no. 3, pp. 199–214, 2005.
- [15] S. Song, Z. Lu, and H. Qiao, "Subset simulation for structural reliability sensitivity analysis," *Reliability Engineering & System Safety*, vol. 94, no. 2, pp. 658–665, 2009.
- [16] W. Zhao, J. K. Liu, and J. J. Ye, "A new method for parameter sensitivity estimation in structural reliability analysis," *Applied Mathematics and Computation*, vol. 217, no. 12, pp. 5298–5306, 2011.
- [17] S.-K. Au and J. L. Beck, "Estimation of small failure probabilities in high dimensions by Subset Simulation," *Probabilistic Engineering Mechanics*, vol. 16, no. 4, pp. 263–277, 2001.
- [18] F. Miao and M. Ghosn, "Modified subset simulation method for reliability analysis of structural systems," *Structural Safety*, vol. 33, no. 4, pp. 251–260, 2011.
- [19] I. Papaioannou, W. Betz, K. Zwirgmaier, and D. Straub, "MCMC algorithms for subset simulation," *Probabilistic Engineering Mechanics*, vol. 41, pp. 89–103, 2015.
- [20] R. Wang, "Comparison of simulation methods applied to steel bridge reliability evaluations," in *Proceedings of the 13th International Conference on Applications of Statistics and Probability in Civil Engineering, ICASP13*, Seoul, South Korea, May 2019.
- [21] R. M. Neal, "MCMC using Hamiltonian Dynamics," *Handbook of Markov Chain Monte Carlo*, vol. 2, pp. 113–162, 2011.
- [22] S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth, "Hybrid monte carlo," *Physics Letters B*, vol. 195, no. 2, pp. 216–222, 1987.
- [23] R. M. Neal, "Bayesian learning for neural networks," *Lecture Notes in Statistics*, Springer-Verlag, no. 118, Berlin, Germany, 1996.
- [24] E. Akhmatkaya and S. Reich, "GSHMC: an efficient method for molecular simulations," *Journal of Computational Physics*, vol. 227, pp. 4934–4954, 2008.
- [25] H. Strathmann, D. Sejdinovic, S. Livingstone, Z. Szabo, and A. Gretton, "Gradient free Hamiltonian Monte Carlo with efficient kernel exponential families," *Advances in Neural Information Processing Systems*, vol. 30, pp. 955–963, 2015.
- [26] S. H. Cheung and J. L. Beck, "Bayesian model updating using Hybrid Monte Carlo simulation with application to structural dynamic models with many uncertain parameters," *Journal of Engineering Mechanics*, vol. 135, no. 4, pp. 243–255, 2009.
- [27] J. B. Nagel and B. Sudret, "Hamiltonian Monte Carlo and borrowing strength in hierarchical inverse problems," *ASCE-ASME J Risk Uncertainty Eng Syst Part A: Civil Engineering*, vol. 2, no. 3, Article ID B4015008, 2016.
- [28] Z. Wang, M. Broccardo, and J. Song, "Hamiltonian Monte Carlo methods for Subset Simulation in reliability analysis," *Structural Safety*, vol. 79, pp. 51–67, 2019.
- [29] T. J. O'Grady II, D. T. Hisey, and J. F. Kiefner, "Pressure calculation for corroded pipe developed," *Oil & Gas Journal*, vol. 90, p. 442, 1992.
- [30] A. K. Sheikh and D. A. Hansen, "Statistical modelling of pitting corrosion and pipeline reliability," *Corrosion Science*, vol. 46, no. 3, pp. 190–197, 1996.
- [31] R. Xu, J. Zhang, and S. U. Cheng, "Time-domain explicit formulation subset simulation method for dynamic reliability of structures subjected to nonstationary random excitations," *Engineering Mechanics*, vol. 30, no. 7, pp. 28–33, 2013.
- [32] S. K. Au and J. L. Beck, "Estimation of small failure probabilities in high dimensions by subset simulation," *Probabilistic Engineering Mechanics*, vol. 16, pp. 263–277, 2001.
- [33] R. M. Neal, "MCMC using hamiltonian dynamics," *Handbook of Markov Chain Monte Carlo*, vol. 2, pp. 113–162, 2011, <https://arxiv.org/abs/1206.1901>.
- [34] S.-K. Au, "On MCMC algorithm for subset simulation," *Probabilistic Engineering Mechanics*, vol. 43, pp. 117–120, 2016.
- [35] B. Leimkuhler and S. Reich, *Simulating Hamiltonian Dynamics*, Cambridge University Press, Cambridge, UK, 2004.
- [36] M. D. Hoffman and A. Gelman, "The No-U-Turn Sampler: adaptively setting path lengths in Hamiltonian Monte Carlo," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1593–1623, 2014.
- [37] P. Thoft-Christensen and M. J. Baker, *Structural Reliability Theory and its Applications*, Springer-Verlag, Berlin, Germany, 1982.
- [38] H. A. Jensen, F. Mayorga, and M. A. Valdebenito, "Reliability sensitivity estimation of nonlinear structural systems under stochastic excitation: a simulation-based approach," *Computer Methods in Applied Mechanics and Engineering*, vol. 289, pp. 1–23, 2015.

Research Article

Fault Diagnosis of Rolling Bearing Based on Improved VMD and KNN

Quanbo Lu ¹, Xinqi Shen,² Xiujun Wang,³ Mei Li ¹, Jia Li,¹ and Mengzhou Zhang¹

¹College of Information Engineering, China University of Geosciences, Beijing 100083, China

²China Academy of Information and Communications Technology, Beijing 100191, China

³Smart Chip Testing Department, Beijing Software Testing & QA Center, Beijing 100193, China

Correspondence should be addressed to Mei Li; maggieli@cugb.edu.cn

Received 7 July 2021; Revised 24 August 2021; Accepted 16 September 2021; Published 4 October 2021

Academic Editor: Jie Chen

Copyright © 2021 Quanbo Lu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Variational modal decomposition (VMD) has the end effect, which makes it difficult to efficiently obtain fault eigenvalues from rolling bearing fault signals. Inspired by the mirror extension, an improved VMD is proposed. This method combines VMD and mirror extension. The mirror extension is a basic algorithm to inhibit the end effect. A comparison is made with empirical mode decomposition (EMD) for fault diagnosis. Experiments show that the improved VMD outperforms EMD in extracting the fault eigenvalues. The performance of the new algorithm is proven to be effective in real-life mechanical fault diagnosis. Furthermore, in this article, combining with singular value decomposition (SVD), fault eigenvalues are extracted. In this way, fault classification is realized by K-nearest neighbor (KNN). Compared with EMD, the proposed approach has advantages in the recognition rate, which can accurately identify fault types.

1. Introduction

VMD is a novel method to deal with the signals. At present, VMD has shown great achievements in the processing of vibration signal, biological signal, and electrical signal. It has been used for detection, diagnosis, and prediction. In the heart sound signal, Babu et al. [1] adopted VMD to extract the Shannon entropy envelope of the heart beat pulse. The approach can accurately identify the signal features reflecting cardiac abnormalities in ECG. Li et al. [2] proposed an adaptive denoising method based on VMD, which is applied to water supply pipeline leakage location. Compared with EMD, VMD can effectively eliminate modal aliasing and is robust to noise and sampling [3, 4].

At present, many scholars use different methods to study the vibration signal of fault bearing [5–11]. In the field of fault diagnosis, some scholars study and apply VMD. Zhang et al. [12] constructed a new technique, which applies VMD to mechanical fault diagnosis. Through failure mechanism analysis, they established the fault model in fast Fourier transform (FFT) and envelope analysis. This approach can

successfully diagnose the fault of rolling bearing. Jiang et al. [13] presented an adaptive detection method, which uses VMD to detect early defects in bearings. Yi et al. [14] employed VMD to extract bearing fault characteristics. Lv et al. [15] used VMD and multicore vector to realize mechanical fault diagnosis. Compared with the traditional fault diagnosis model, this approach had better performance. Yan et al. [16] adopted VMD to extract hybrid-domain features, identifying fault type with high accuracy. VMD has an important application value for mechanical fault diagnosis [17]. However, the drawback of the abovementioned approaches is that they have the modal aliasing in VMD, which affects the diagnostic performance.

In this paper, a new method is proposed for rolling bearing fault diagnosis. First, a mirror extension to suppress its modal aliasing is used and the improved VMD into bearing fault diagnosis is introduced [18]. Then, combine the improved VMD with SVD to achieve the effective extraction of fault eigenvalues, finally adopt KNN to complete fault classification. The contributions of this paper are presented as follows:

- (1) An improved VMD approach is proposed based on VMD and mirror extension, which inhibits the end effect. Improved VMD makes it effective for fault diagnosis.
- (2) Compared with EMD, the proposed approach has an advantage in the recognition rate, which can accurately identify the rolling bearing fault type.
- (3) The proposed diagnosis framework is effectively verified by detection of the rolling bearing fault. Results demonstrate that our framework has advantages in extracting the characteristic frequency, which is suitable for the detection of manufacturing systems.

The article is organized as follows. The principles of VMD, mirror extension, and KNN are introduced in Section 2. Section 3 provides a new method and the detailed diagnosis scheme of the proposed approach. Section 4 presents the experimental verification. Section 5 concludes this paper.

2. Theories

2.1. VMD. VMD is a new approach of signal processing, which is adaptive and quasi-orthogonal. The original signal

is decomposed into several modal components using VMD. Each modal component has limited bandwidth, and it is near the central frequency w_k [19]. As presented in equation (1), the constrained variational models are calculated:

$$\min_{\{u_k\}, \{w_k\}} \left\{ \sum_k \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) u_k(t) \right] e^{-jw_k t} \right\|_2^2 \right\} \quad (1)$$

s.t. $\sum_k u_k = f,$

where $\{u_k\} = \{u_1, \dots, u_k\}$ are the narrow-band components and $\{w_k\} = \{w_1, \dots, w_k\}$ represent the central frequencies. $\sum_k = \sum_{k=1}$, where k is the number of decompositions.

In order to find the optimal solution of the constrained variational model, the VMD method uses the quadratic penalty function with good convergence and the Lagrange multiplier λ operator with strong constraint ability. Therefore, a Lagrangian function L can be introduced to optimize the constrained variational problem so as to minimize the narrow-band component u_k and the central frequency w_k obtained in equations (3) and (4). The expression of L is introduced:

$$L(\{u_k\}, \{w_k\}, \lambda) = \alpha \sum_k \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) u_k(t) \right] e^{-jw_k t} \right\|_2^2 + \left\| f(t) - \sum_k u_k(t) \right\|_2^2 + \langle \lambda(t), f(t) - \sum_k u_k(t) \rangle. \quad (2)$$

The formula for minimizing the narrow-band component \hat{u}^k is expressed as follows:

$$\hat{u}_k^{n+1}(w) = \frac{\hat{f}(w) - \sum_{i \neq k} \hat{u}_i(w) + \hat{\lambda}(w)/2}{1 + 2\alpha(w - w_k)^2}. \quad (3)$$

The formula for minimizing the central frequency narrow-band component u_k is obtained as follows:

$$w_k^{n+1} = \frac{\int_0^\infty w |\hat{u}(w)|^2 dw}{\int_0^\infty |\hat{u}(w)|^2 dw}. \quad (4)$$

2.2. KNN. KNN is a nonparametric prediction algorithm. It searches for k most similar eigenvectors in the historical database to predict the future value [20]. The model has simple structure and high computational efficiency.

The KNN classification algorithm is described as follows. When the testing sample (unknown sample) is given, first search the pattern space to find the K training sample closest to the testing sample, that is, K -nearest neighbor, and then calculate the selected K -nearest neighbor. If a class has the maximum number of the nearest neighbors, the testing sample can be determined to the class. Euclidean distance is used to calculate the distance between the testing sample and all training samples. The formula is defined as follows:

$$d(X, Y) = \sqrt{\sum_{i=1}^N (x_i - y_i)^2}, \quad (5)$$

where X is the testing sample and Y denotes the training sample. Figure 1 is the detailed steps of KNN.

3. The Proposed Diagnosis Method

3.1. Improved VMD. The mirror extension assumes that a mirror is placed at both ends of the data. The original data sequence image in the mirror is symmetrical with respect to the mirror and the original data sequence [21]. The original data sequence image and the original data sequence in the two mirrors form a continuous and closed ring without endpoints. The internal data completely determines the upper and lower envelopes of the data, so the end effect is fundamentally avoided [22]. Above the mirror surface is the original data. Below the mirror surface is the extension data. After the operation, only the data processing results above the mirror surface are output.

The purpose of boundary extension is not to provide accurate data beyond the end, but to provide a condition that the data within the end completely determine the envelope. The mirror extension satisfies this condition, so it is an ideal extension algorithm.

When VMD analyzes and processes the fault signal, the error is caused by the influence of external factors on both ends [23–27]. Taking into account the end effect in VMD, a mirror extension is adopted to suppress it. The newly proposed method has an advantage in extracting the characteristic frequency. Figure 2 is the detailed steps of the

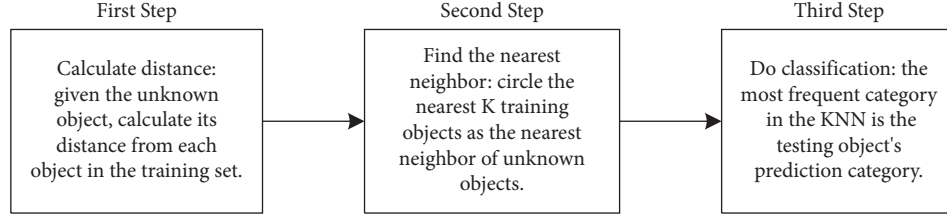


FIGURE 1: Steps of KNN.

improved VMD. The main process of improved VMD can be summarized as follows:

Step 1 : initialize $\{\hat{u}_k^1\}, \{w_k^1\}, \hat{\lambda}^1, n \leftarrow 0$, where $\{u_k\} = \{u_1, \dots, u_k\}$ are the narrow-band components, $\{w_k\} = \{w_1, \dots, w_k\}$ represent the central frequencies, k is the number of decompositions, and λ represents the Lagrange multiplier.

Step 2 : set $n \leftarrow n + 1$ and $k \leftarrow k + 1$ and execute the whole cycle. Update \hat{u}_k and w_k for all $w \geq 0$ to reach the preset decomposition number. When $k = K$, the cycle ends. The updated formula of the narrow-band component and the corresponding central frequency are as follows:

$$\hat{u}_k^{n+1}(w) = \frac{\hat{f}(w) - \sum_{i < k} \hat{u}_i^{n+1}(w) - \sum_{i > k} \hat{u}_i^n(w) + \hat{\lambda}^n(w)/2}{1 + 2a(w - w_k^n)^2}, \quad (6)$$

$$\hat{w}_k^{n+1} = \frac{\int_0^\infty w |\hat{u}_k^{n+1}(w)|^2 dw}{\int_0^\infty |\hat{u}_k^{n+1}(w)|^2 dw}. \quad (7)$$

Step 3 : update λ according to the formula:

$$\hat{\lambda}^{n+1}(w) \leftarrow \hat{\lambda}^n(w) + \tau \left(\hat{f}(w) - \sum_k \hat{u}_k^{n+1}(w) \right). \quad (8)$$

Step 4 : return to Step 2 and repeat the above process until the whole iterative process meets the constraints, and a series of narrow-band eigenmode component signals are obtained. Equation (9) is the constraint condition, where ε is set to 10^{-6} :

$$\frac{\sum_k \|\hat{u}_k^{n+1} - \hat{u}_k^n\|_2^2}{\|\hat{u}_k^n\|_2^2} < \varepsilon. \quad (9)$$

The specific extension process is provided as follows:

(1) Find all extreme points of the rolling bearing fault signal $x(t)$, $t = 1, 2, \dots, T$. The extreme points include the local maximum points and the local minimum points.

The local maximum point sequence of the signal $x(t)$ is

$$\{(f_{\max}(1), x_{\max}(1)), (f_{\max}(2), x_{\max}(2)), \Lambda(f_{\max}(t_{\max}), x_{\max}(t_{\max}))\}. \quad (10)$$

The local minimum point sequence of the signal $x(t)$ is

$$\{(f_{\min}(1), x_{\min}(1)), (f_{\min}(2), x_{\min}(2)) \Lambda(f_{\min}(t_{\min}), x_{\min}(t_{\min}))\}, \quad (11)$$

where $f_{\max}(t)$, $t = 1, 2, \dots, T$, are the local maximum points, $f_{\min}(t)$, $t = 1, 2, \dots, T$, are the local minimum points, $x_{\max}(t)$, $t = 1, 2, \dots, T$, are the maximum points, and $x_{\min}(t)$, $t = 1, 2, \dots, T$, are the minimum points.

(2) According to equation (12), the symmetrical extension is employed to the extreme points of the fault signal $x(t)$ to obtain a new extreme point sequence $\{p'\}$:

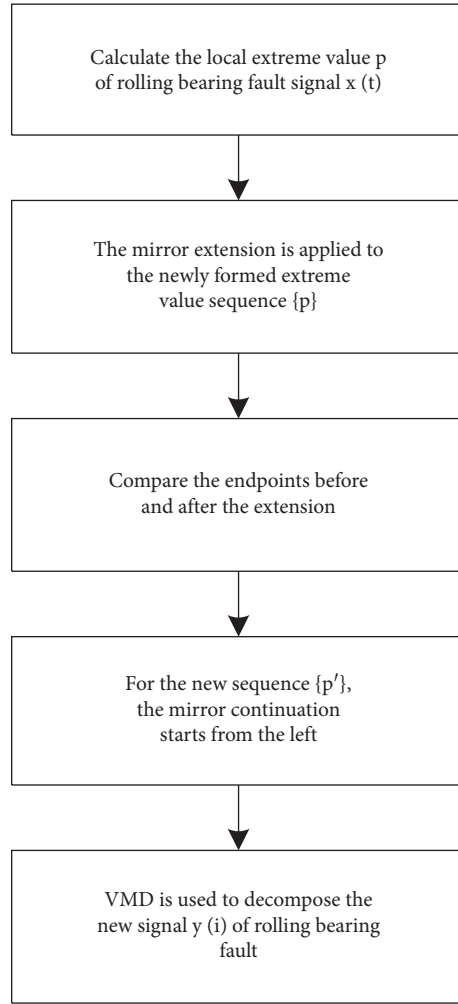


FIGURE 2: Steps of the improved VMD.

$$\left\{ \begin{array}{l} \left\{ \begin{array}{l} f_{\max}(0) = f_{\max}(1) \\ x_{\max}(0) = 2 * x_{\max}(1) - x_{\max}(2), \text{ when } x_{\min}(1) < x_{\max}(1), \end{array} \right. \\ \left\{ \begin{array}{l} f_{\min}(0) = f_{\min}(1) \\ x_{\min}(0) = 2 * x_{\min}(1) - x_{\min}(2), \text{ when } x_{\min}(1) > x_{\max}(1), \end{array} \right. \\ \left\{ \begin{array}{l} f_{\max}(t_{\max} + 1) = f_{\max}(t_{\max}) \\ x_{\max}(t_{\max} + 1) = 2 * x_{\max}(t_{\max}) - x_{\max}(t_{\max} - 1), \text{ when } x_{\min}(t_{\min}) > x_{\max}(t_{\max}) \end{array} \right. \\ \left\{ \begin{array}{l} f_{\min}(t_{\min} + 1) = f_{\min}(t_{\min}) \\ x_{\min}(t_{\min} + 1) = 2 * x_{\min}(t_{\min}) - x_{\min}(t_{\min} - 1), \text{ when } x_{\min}(t_{\min}) < x_{\max}(t_{\max}). \end{array} \right. \end{array} \right. \quad (12)$$

(3) Correct the new extreme point sequence $\{p'\}$ after extension:

$$\left\{ \begin{array}{l} \left\{ \begin{array}{l} f_{\min}(0) > x(1), \text{ when } x_{\max}(1) < x_{\min}(1) \\ f_{\max}(0) > x(1), \text{ when } x_{\max}(1) > x_{\min}(1) \end{array} \right. \left\{ \begin{array}{l} f_{\min}(t_{\min}) > x(t), \text{ when } x_{\min}(t_{\min}) < x_{\max}(t_{\max}) \\ f_{\max}(t_{\max}) < x(t), \text{ when } x_{\min}(t_{\min}) > x_{\max}(t_{\max}) \end{array} \right. \end{array} \right. \quad (13)$$

(4) Use the value on the left end of the fault signal $x(t)$ as the symmetry plane, and the mirror extension is

realized. Meanwhile, the new fault signal $y(i)$ of rolling bearing is realized.

3.2. Diagnosis Scheme. Owing to the feature extraction is difficult in the incipient rolling bearing failures, this paper presents a new method combined improved VMD and KNN. The structure of the diagnosis scheme is shown in Figure 3.

The detailed steps of the diagnosis scheme are as follows:

- (1) Firstly, improved VMD is applied to handle the fault signal, and several narrow-band eigenmode components u are realized.
- (2) Then, the narrow-band components are selected, which have rich fault information. These modal components are selected to reconstruct the fault signal, and the envelope spectrum is conducted for it.
- (3) Finally, the feature vector from modal components is extracted by using SVD. After feature extraction, there are 100 data sets. 70 of the 100 data sets are randomly employed to train the KNN classification model, and the rest are employed to test the classification performance.

4. Experimental Verification

4.1. Experimental Setting. As shown in Figure 4, the fault diagnosis platform consists of a 0.75 kW three-phase asynchronous motor, two couplings, a reducer, a magnetic powder brake, a piezoelectric accelerometer, a faulty rolling bearing, and a photoelectric speed sensor. The faulty rolling bearing is installed in the bearing pedestal of the reducer.

The signal acquisition system is used to collect the fault signals, in which the rotation speed n and the sampling frequency f are set at 600 r/min and 1 kHz, respectively. Table 1 is the basic parameters of rolling bearing. According to these parameters, the fault characteristic frequencies are calculated: inner race is 99 Hz and outer race is 71 Hz. The improved VMD is applied to decompose the fault signal. When the mode number K is different, their central frequency is different. The relationship between them is depicted in Figure 5.

4.2. Experimental Results and Analysis. When the value of K starts from 5, the central frequency is close [26], see Figure 5(a). This is an over decomposition phenomenon. Hence, the K value taken in the test is 5. Based on VMD experience, the balance parameter constrained by data fidelity adopts the default value of 2000, and the time step of the double rise is 0.1. Figure 6(a) illustrates the time domain diagram of the inner race fault, and its improved VMD result is demonstrated in Figure 7(a).

As illustrated in Figure 5(b), when the value of K starts from 5, the central frequency is close, which is an over decomposition phenomenon [28]. Thereby, the K value taken in the test is 5. Based on VMD experience, the balance parameter constrained by data fidelity adopts the default value of 2000, and the time step of the double rise is 0.1. Figure 6(b) shows the time domain diagram of the outer race fault. The improved VMD result is demonstrated in Figure 7(b).

As shown in Figure 6(a), the original signal of the inner race fault contains considerable background noise, which obscures the fault information. As shown in Figure 7(a), compared with the original signal, the modal components after the improved VMD eliminate the noise, which is the function of Wiener filter. There are obvious fault shock components in U_3 , U_4 , and U_5 . Therefore, these three modal components are selected to reconstruct the signal, and the envelope spectrum is conducted for it. The result of the envelope spectrum is depicted in Figure 8(a).

In contrast to the improved VMD, the first five components of EMD are also taken, and Figure 9(a) is the decomposition results. In Figure 9(a), IMF_1 , IMF_2 , and IMF_3 contain many fault feature information, and they are reconstructed. Figure 8(b) presents the envelope spectrum of the reconstructed signal.

From Figure 6(b), the original signal time domain of the outer race fault contains large background noise, which obscures the fault information. Figure 7(b) shows that the modal components eliminate the noise, which is the function of the Wiener filter. There are obvious fault shock components in U_2 , U_4 , and U_5 . Hence, these three modal components are selected to reconstruct the signal, and the envelope spectrum is conducted for it. The result of the envelope spectrum is provided in Figure 10(a).

As a comparative test analysis, the first five components of EMD are taken, and Figure 9(b) is the decomposition result. As evident in Figure 9(b), IMF_1 , IMF_2 , and IMF_3 contain many fault feature information, and they are reconstructed. Figure 10(b) presents the envelope spectrum of the reconstructed signal.

Figure 8(a) displays that the envelope spectrum has obvious fault shock characteristics around 99 Hz, which is basically consistent with the theoretical value. The small gap does not affect the fault identification. 198 Hz is the second frequency, and 297 Hz is the third frequency. As shown in Figure 8(b), the envelope spectrum is not ideal. 99 Hz and its octave are not demodulated. Consequently, the decomposition result of the improved VMD outperforms EMD.

Figure 10(a) shows that the envelope spectrum has obvious fault shock characteristics around 71 Hz. This is basically consistent with the theoretical value. The small gap does not affect the fault identification. 142 Hz is the second frequency, and 213 Hz is the third frequency. In Figure 10(b), the envelope spectrum is not ideal. 71 Hz and its octave are not demodulated. Therefore, the decomposition result of the improved VMD is shown to outperform EMD.

In the fault classification, K in the KNN algorithm is set to 1. The first three groups of singular values are selected as the nearest neighbor classification training sets and testing sets. For each fault type, 70 of the 100 datasets are randomly employed to train the KNN classification model, and the rest are applied to test the classification performance. The fault classification results after improved VMD + SVD are shown in Figure 11(a), and the fault classification results after EMD + SVD are provided in Figure 11(b). Figures 11(a) and 11(b) show that the performance of the improved VMD + SVD is better than EMD + SVD.

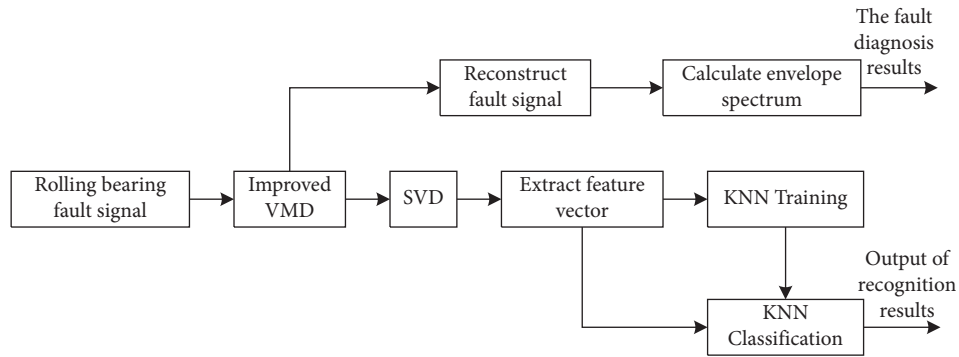


FIGURE 3: The structure of the diagnosis scheme.

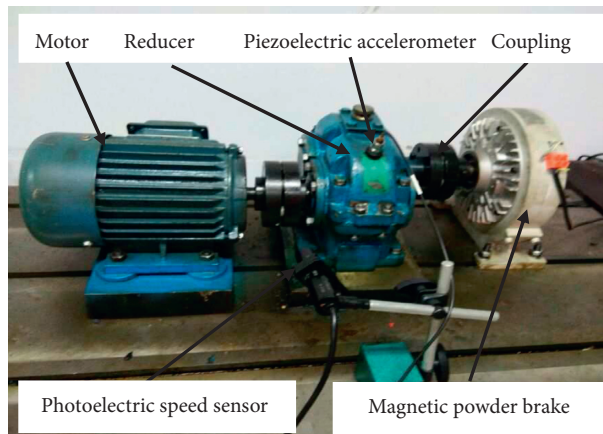


FIGURE 4: The fault diagnosis platform.

TABLE 1: The parameters of bearing.

The diameter of angle rolling elements	Bearing pitch	The number of rolling elements	Contact angle
8	46	17	14°

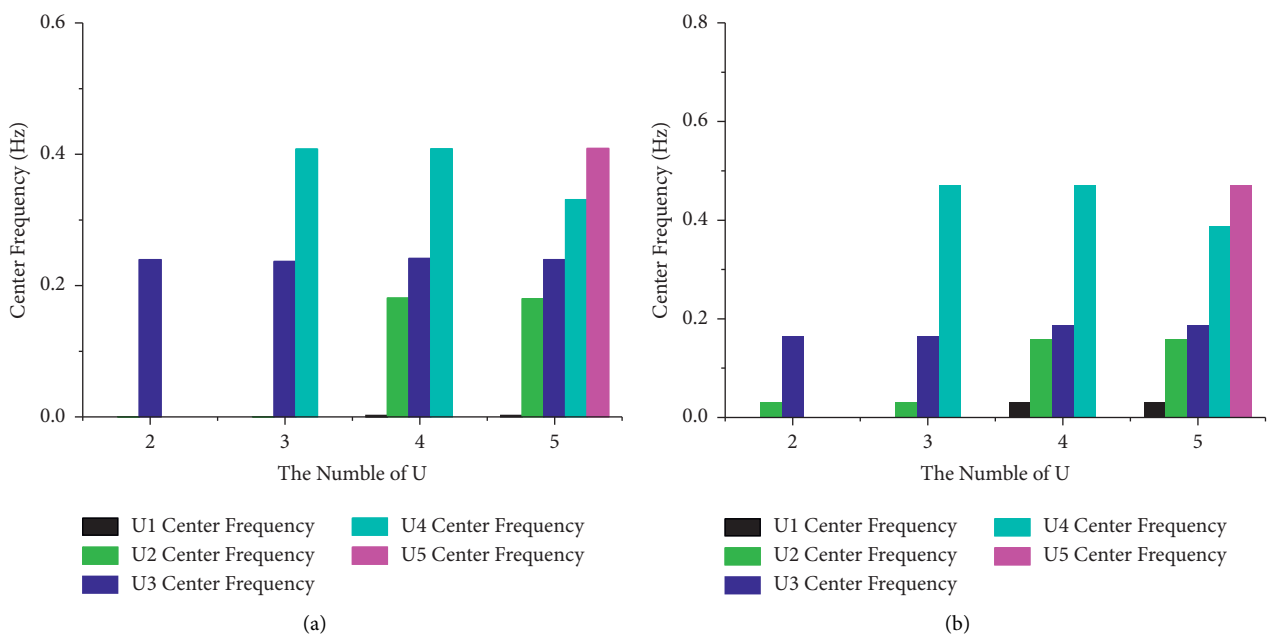


FIGURE 5: The relationship between mode number and central frequency after improved VMD. (a) Inner race fault. (b) Outer race fault.

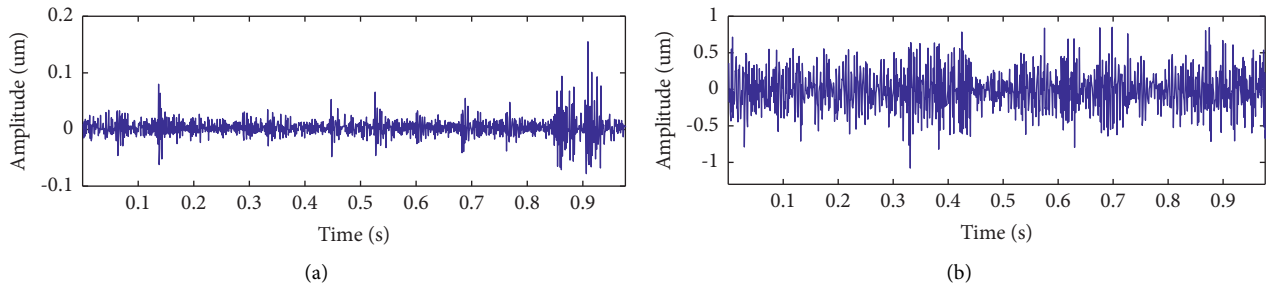


FIGURE 6: Time domain diagram. (a) Inner race fault. (b) Outer race fault.

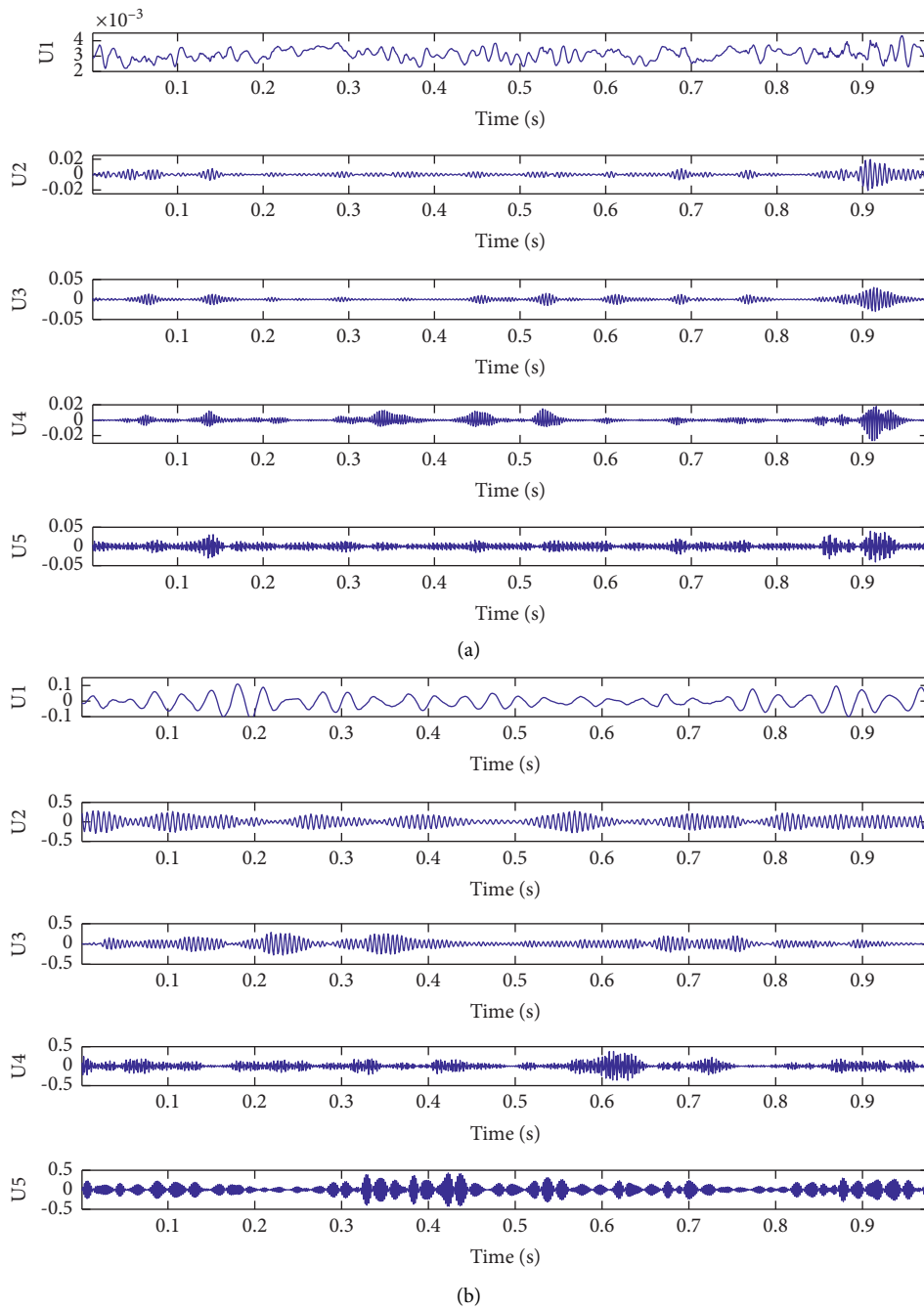


FIGURE 7: The decomposition results after improved VMD. (a) Inner race fault. (b) Outer race fault.

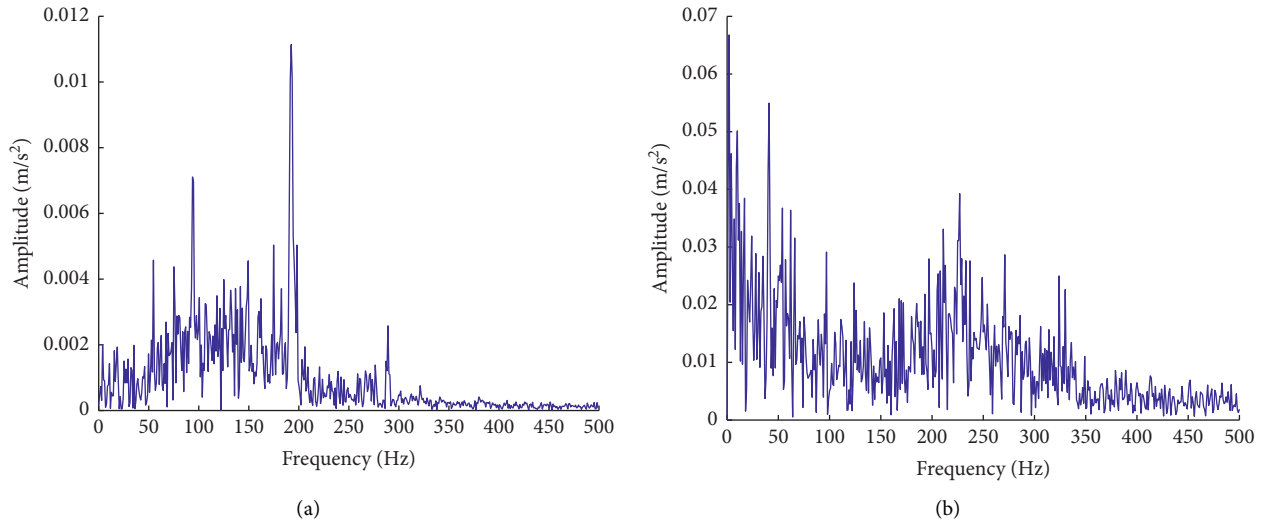
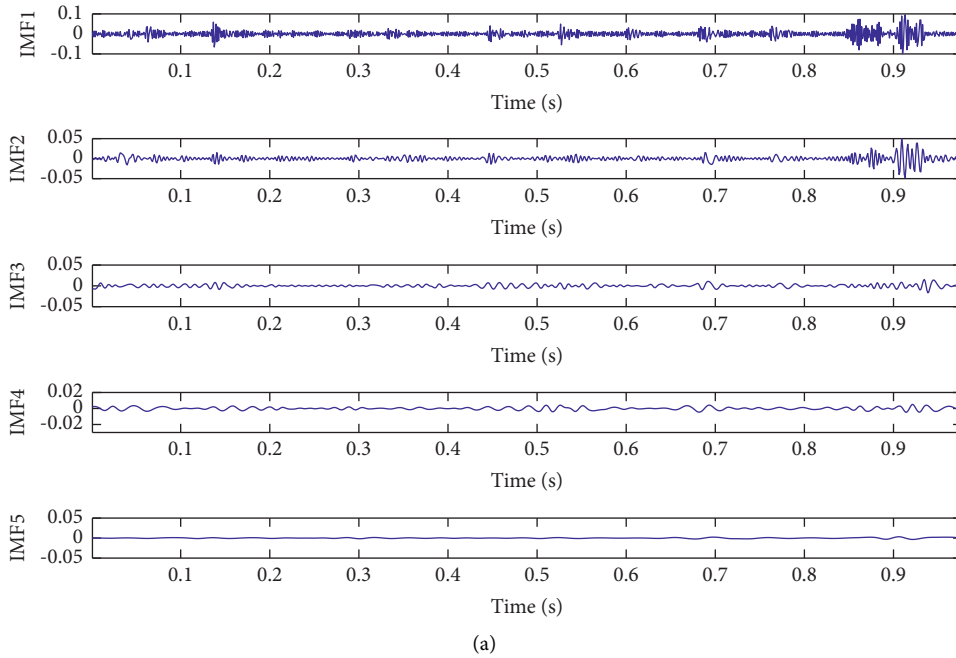


FIGURE 8: Envelope spectrum of inner race fault. (a) $U_3 + U_4 + U_5$. (b) $IMF_1 + IMF_2 + IMF_3$.



(a)
FIGURE 9: Continued.

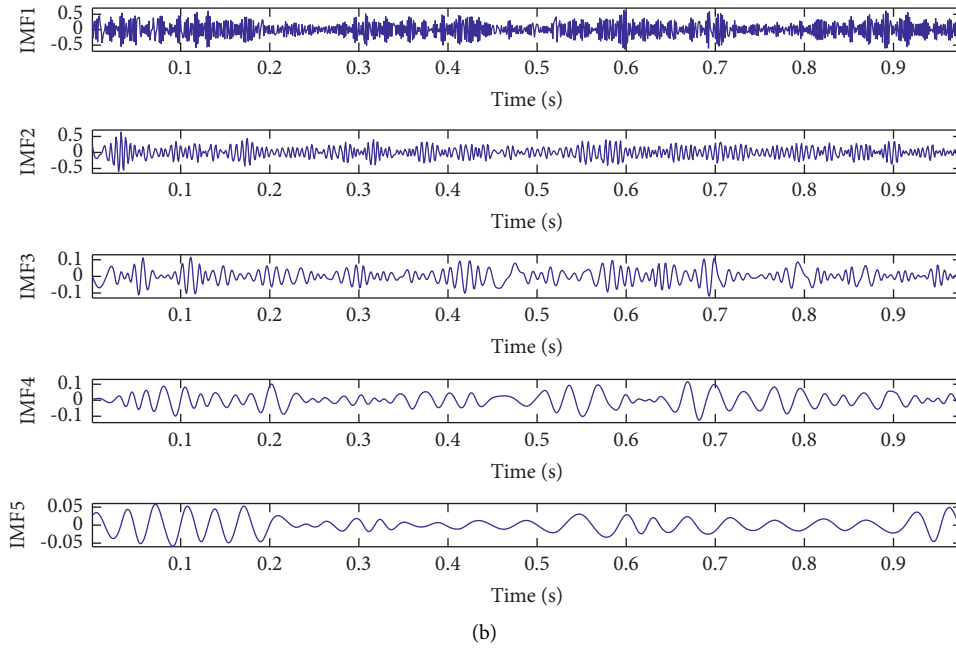


FIGURE 9: The decomposition results after EMD. (a) Inner race fault. (b) Outer race fault.

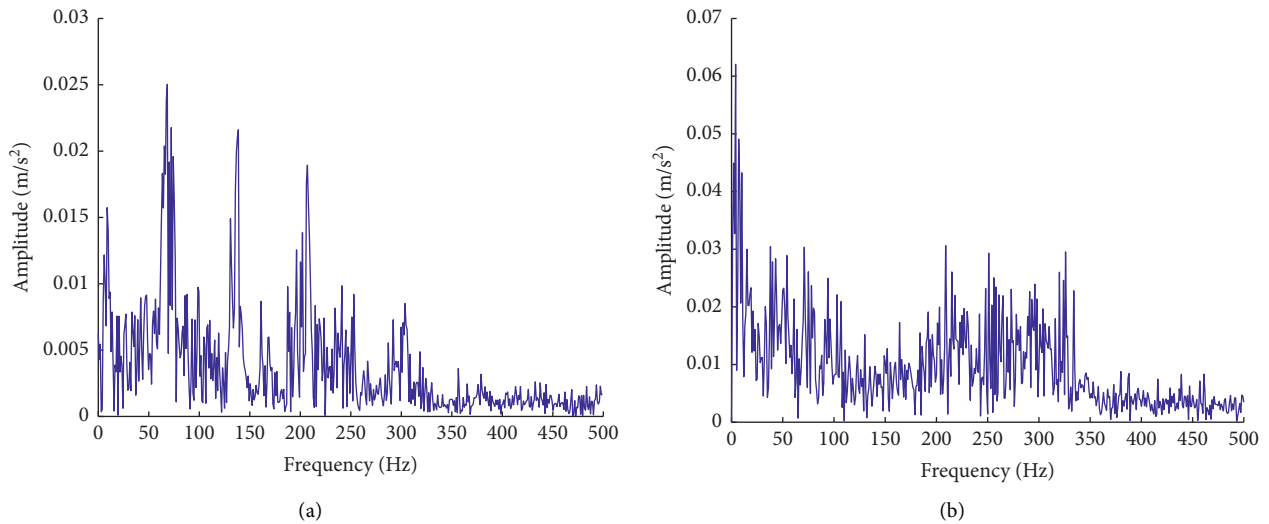


FIGURE 10: Envelope spectrum of outer race fault. (a) $U_2 + U_4 + U_5$. (b) $IMF_1 + IMF_2 + IMF_3$.

The fault recognition rates are reported in Table 2. In the classification results after the improved VMD + SVD, one inner race fault is incorrectly identified as the outer race fault, and one outer race fault is incorrectly identified as the

inner race fault. In the classification results after EMD + SVD, three inner race faults are incorrectly identified as the outer race faults, and five outer race faults are incorrectly identified as the inner race faults.

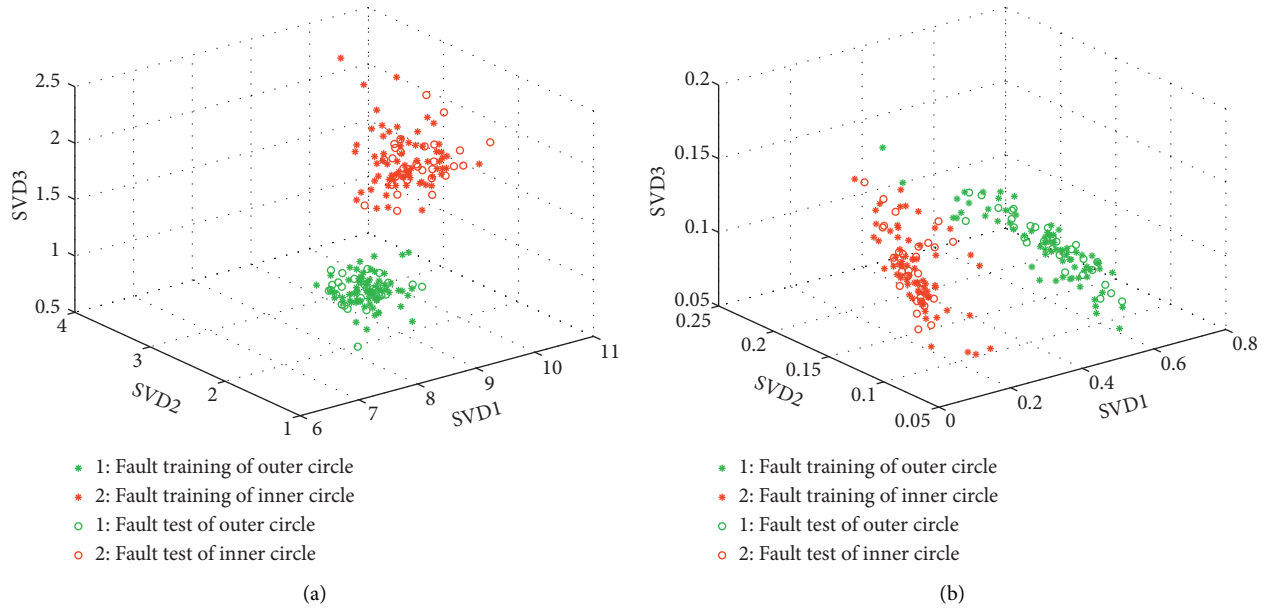


FIGURE 11: Fault classification result. (a) Improved VMD + SVD and (b) EMD + SVD.

TABLE 2: The fault recognition rates.

Algorithm	Inner race fault	Outer race fault	Recognition rate (%)
Improved VMD + SVD	29	29	96.67
EMD + SVD	27	25	86.67

5. Conclusion

A new method combining the improved VMD and KNN for the rolling bearing fault detection system is presented. First, the improved VMD decomposes the fault signal to obtain the modal component. Then, the modal components with many fault shock components are reconstructed. Finally, the envelope spectrum is applied to the reconstructed signal, and the characteristic frequency is identified successfully. Furthermore, combining with SVD, KNN realizes fault type classification. The experimental results demonstrate that the proposed approach has potential application in rolling bearing fault diagnosis.

Data Availability

Some or all data, models, or code generated or used during the study are available from the corresponding author upon request.

Conflicts of Interest

The authors have no conflicts of interest/competing interest to disclose.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant no. 41572347).

References

- [1] K. A. Babu, B. Ramkumar, and M. S. Manikandan, "Automatic identification of S1 and S2 heart sounds using simultaneous PCG and PPG recordings," *IEEE Sensors Journal*, vol. 18, no. 22, pp. 9430–9440, 2018.
- [2] J. Li, Y. Chen, Z. Qian, and C. Lu, "Research on VMD based adaptive denoising method applied to water supply pipeline leakage location," *Measurement*, vol. 151, pp. 1–13, 2020.
- [3] A. Dibaj, M. Mohammad Etefagh, R. Hassannejad, and M. Biuok Ehghaghi, "A hybrid fine-tuned VMD and CNN scheme for untrained compound fault diagnosis of rotating machinery with unequal-severity faults," *Expert Systems with Applications*, vol. 11, pp. 1–16, 2020.
- [4] A. Joshuva, R. S. Kumar, S. Sivakumar, G. Deenadayalan, and R. Vishnuvardhan, "An insight on VMD for diagnosing wind turbine blade faults using C4.5 as feature selection and discriminating through multilayer perceptron," *Alexandria Engineering Journal*, vol. 59, no. 5, pp. 3863–3879, 2020.
- [5] W. Deng, H. Liu, J. Xu, H. Zhao, and Y. Song, "An improved quantum-inspired differential evolution algorithm for deep belief network," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 10, pp. 7319–7327, 2020.
- [6] H. M. Zhao, S. Y. Zuo, M. Hou et al., "A novel adaptive signal processing method based on enhanced empirical wavelet transform technology," *Sensors*, vol. 18, pp. 1–17, 2018.
- [7] M. Kuncan, "An intelligent approach for bearing fault diagnosis: combination of 1D-LBP and GRA," *IEEE Access*, vol. 8, pp. 137517–137529, 2020.
- [8] K. Zhang, Y. Xu, Z. Liao, L. Song, and P. Chen, "A novel fast entrogram and its applications in rolling bearing fault

- diagnosis,” *Mechanical Systems and Signal Processing*, vol. 154, Article ID 107582, 2021.
- [9] J. Cheng, Y. Yang, X. Li, and J. Cheng, “Adaptive periodic mode decomposition and its application in rolling bearing fault diagnosis,” *Mechanical Systems and Signal Processing*, vol. 161, Article ID 107943, 2021.
- [10] X. Wang, D. Mao, and X. Li, “Bearing fault diagnosis based on vibro-acoustic data fusion and 1D-CNN network,” *Measurement*, vol. 173, Article ID 108518, 2021.
- [11] J. Wei, H. Huang, L. Yao, Y. Hu, Q. Fan, and D. Huang, “New imbalanced bearing fault diagnosis method based on Sample-characteristic oversampling technique (SCOTE) and multi-class LS-SVM,” *Applied Soft Computing*, vol. 101, Article ID 107043, 2021.
- [12] M. Zhang, Z. Jiang, and K. Feng, “Research on variational mode decomposition in rolling bearings fault diagnosis of the multistage centrifugal pump,” *Mechanical Systems and Signal Processing*, vol. 93, pp. 460–493, 2017.
- [13] X. Jiang, S. Li, and C. Cheng, “A novel method for adaptive multiresonance bands detection based on VMD and using MTEO to enhance rolling element bearing fault diagnosis,” *Shock and Vibration*, vol. 2016, Article ID 8361289, 20 pages, 2016.
- [14] C. Yi, Y. Lv, and Z. Dang, “A fault diagnosis scheme for rolling bearing based on particle swarm optimization in variational mode decomposition,” *Shock and Vibration*, vol. 2016, Article ID 9372691, 10 pages, 2016.
- [15] Z. Lv, B. Tang, Y. Zhou, and C. Zhou, “A novel method for mechanical fault diagnosis based on variational mode decomposition and multikernel support vector machine,” *Shock and Vibration*, vol. 2016, 11 pages, 2016.
- [16] X. Yan, Y. Liu, M. Jia, and Y. Zhu, “A multi-stage hybrid fault diagnosis approach for rolling element bearing under various working conditions,” *IEEE Access*, vol. 7, pp. 138426–138441, 2019.
- [17] X. An and J. Yang, “Denoising of hydropower unit vibration signal based on variational mode decomposition and approximate entropy,” *Transactions of the Institute of Measurement and Control*, vol. 38, no. 3, pp. 282–292, 2016.
- [18] F. Jiang, Z. Zhu, and W. Li, “An improved VMD with empirical mode decomposition and its application in incipient fault detection of rolling bearing,” *IEEE Access*, vol. 6, pp. 44483–44493, 2018.
- [19] K. Dragomiretskiy and D. Zosso, “Variational mode decomposition,” *IEEE Transactions on Signal Processing*, vol. 62, no. 3, pp. 531–544, 2014.
- [20] N. Ling, G. Aneiros, and P. Vieu, “kNN estimation in functional partial linear modeling,” *Statistical Papers*, vol. 61, no. 1, pp. 423–444, 2020.
- [21] C. Dong, X. Jiao, and F. Xu, “Mirror extensions of vertex operator algebras,” *Communications in Mathematical Physics*, vol. 329, no. 1, pp. 263–294, 2014.
- [22] J. Wang, W. Liu, and S. Zhang, “An approach to eliminating end effects of EMD through mirror extension coupled with support vector machine method,” *Personal and Ubiquitous Computing*, vol. 23, no. 3-4, pp. 443–452, 2019.
- [23] M. G. A. Nassef, T. M. Hussein, and O. Mokhiamar, “An adaptive variational mode decomposition based on sailfish optimization algorithm and gini index for fault identification in rolling bearings,” *Measurement*, vol. 108514, pp. 1–20, 2020.
- [24] S. Braut, R. Žigulić, A. Skoblar, and G. Štimac Rončević, “Partial rub detection based on instantaneous angular speed measurement and variational mode decomposition,” *Journal of Vibration Engineering & Technologies*, vol. 8, no. 2, pp. 351–364, 2020.
- [25] Z. Wang, J. Wang, and W. Du, “Research on fault diagnosis of gearbox with improved variational mode decomposition,” *Sensors*, vol. 18, no. 10, p. 3510, 2018.
- [26] Y. Miao, M. Zhao, and J. Lin, “Identification of mechanical compound-fault based on the improved parameter-adaptive variational mode decomposition,” *ISA Transactions*, vol. 84, pp. 82–95, 2019.
- [27] Y. X. Wang, L. Yang, and J. W. Xiang, “A hybrid approach to fault diagnosis of roller bearings under variable speed conditions,” *Measurement Science and Technology*, vol. 4, pp. 1587–1601, 2017.
- [28] H. Liu, D. Li, Y. Yuan, S. Zhang, H. Zhao, and W. Deng, “Fault diagnosis for a bearing rolling element using improved VMD and HT,” *Applied Sciences*, vol. 9, no. 7, p. 1439, 2019.

Research Article

Modeling and Design of an Aircraft-Mode Controller for a Fixed-Wing VTOL UAV

Zhao Deng , Liaoni Wu, and Yancheng You

UAV Lab, School of Aerospace Engineering, Xiamen University, Xiamen, China

Correspondence should be addressed to Zhao Deng; 32020160153977@stu.xmu.edu.cn

Received 26 May 2021; Revised 30 August 2021; Accepted 7 September 2021; Published 29 September 2021

Academic Editor: Jie Chen

Copyright © 2021 Zhao Deng et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Vertical takeoff and landing (VTOL) is an essential feature of unmanned aerial vehicles (UAVs). On the one hand, VTOL can expand and enhance the applications of UAVs; yet, on the other hand, it makes the design of control systems for UAVs more complicated. The most challenging demand in designing the control system is to achieve satisfactory response sharpness of fixed-wing UAVs to control commands and ensure that the aircraft mode channels are effectively decoupled. In this work, a six-degree-of-freedom (6-DoF) model with forces and moments is established based on the aerodynamic analysis, which is carried out through computational fluid dynamics (CFD) numerical simulation. The improved proportional derivative (PD) controller based on the extended state observer (ESO) is proposed to design the inner-loop attitude control, which increases the anti-interference ability for internal and external uncertainty of the UAV system. The motion equations of the UAV are established and divided into independent components of longitudinal and lateral motion to design the outer loop control law under minor disturbance conditions. A total energy control system (TECS) for the longitudinal height channel is proposed, which separates speed control and track control. L1 nonlinear path tracking guidance algorithm is used for lateral trajectory tracking so as to improve curve tracking ability and wind resistance. Effectiveness of this approach is proved by actual flight experiment data. Finally, a controller based on angular velocity control is designed to prevent the attitude and head reference system (AHRS) from malfunctioning. Its effectiveness is verified by the response test of the control system.

1. Introduction

The VTOL functions keep the fixed-wing UAVs from relying on runways to takeoff or land and also significantly reduce the risk of accidents during takeoff and landing. Besides, the VTOL function supplies the fixed-wing UAVs, the flexibility to make it easier to perform tasks in mountainous areas, sea, and other scenarios, thereby broadening their application [1]. Furthermore, UAVs with VTOL can quickly implement one-key autonomous takeoff and landing functions, which improves the reliability and safety in takeoff and landing for potential nonprofessional and unskilled users [2, 3]. Hence, the UAVs with VTOL functions have always been the focus of research in the aviation field of Western Europe and the United States [4, 5]. Typical representatives are tail-mounted VD200, tiltrotor “Osprey” V-22, tilt-wing NASA “Greased Lightning,” and fixed-wing HQ-60 from American Latitude Engineering LLC.

The design of controllers for VTOL UAVs is widely concerned in both the industry and the research field. In the recent years, multiple novel methods have been proposed to improve the performance of VTOL UAVs. For example, Özgür Dündar proposed to employ an aerodynamic design steps and sizing of both wing and control surfaces to improve static stability and endurance [6]. Wu et al. employed multiple sensors to design a UAV system for emergency response [7]. Oca et al. presented a longitudinal aircraft dynamics to model the takeoff and landing considering the rolling resistance forces during ground roll through a friction model [8]. Govdali et al. developed a detailed aerodynamic modeling technique along with a fuzzy switching multimodel guidance and control strategy for a UAVs and successfully controlled the aircraft for a full flight envelope from hover to landing [9]. However, to date, the control method and algorithms have not been fully developed to fulfill the needs for various UAVs. More researches

are still needed to improve the performance of VTOL UAVs in specific circumstances.

Based on novel control techniques and algorithms, this paper takes SD-40 UVA as an example and presents the modeling and design of an aircraft mode controller for fixed-wing UAVs with VTOL functions. The SD-40 UVA has a fixed wing with eight horizontal rotors installed on the double-tail brace, as shown in Figure 1.

In this work, a six-degree-of-freedom (6-DoF) model with forces and moments is established. A improved proportional derivative (PD) controller based on the extended state observer (ESO) and a total energy control system (TECS) for the longitudinal height channel are proposed. The motion equations of the UAV are established. L1 nonlinear path tracking guidance algorithm is used to improve curve tracking ability and wind resistance. Finally, a controller based on angular velocity control is designed to prevent the attitude and head reference system (AHRS) from malfunctioning. Effectiveness of these novel approaches is investigated by analyzing actual flight experiment data.

2. SD-40 Motion Mathematical Model

The nonlinear model obtained by the Newton–Euler formulation involves two main right-hand reference coordinate systems [10], as shown in Figure 2. For the geodetic coordinate frame, which is expressed as $O_g x_g y_g z_g$. A North-East-Down (NED) orthogonal coordinate frame is established with its origin at the takeoff point.

$O_b x_b y_b z_b$ is the body-fixed frame (BFF) attached to the center of gravity (CoG) of SD-40, where $O_b x_b$ points out the nose of the SD-40, $O_b y_b$ points out the right wing, and $O_b z_b$ points out the belly.

B_{bg} represents the transformation matrix from the NED frame to the BFF.

$$B_{bg} = \begin{bmatrix} c\theta c\psi & c\theta s\psi & -s\theta \\ s\theta s\phi c\psi - c\phi s\psi & s\theta s\phi s\psi + c\phi c\psi & c\theta s\phi \\ s\theta c\phi c\psi + s\phi s\psi & s\theta c\phi s\psi - s\phi c\psi & c\theta c\phi \end{bmatrix}, \quad (1)$$

where ϕ , θ , and ψ represent the Euler angles and s and c are the shorthand notation of the sine and cosine functions, respectively.

2.1. Linear Motion Equation. According to the momentum theorem, the kinetic equation of mass center can be obtained:

$$\begin{bmatrix} \dot{U} \\ \dot{V} \\ \dot{W} \end{bmatrix} = \frac{1}{m} \begin{bmatrix} F_x \\ F_y \\ F_z \end{bmatrix} - \begin{bmatrix} P \\ Q \\ R \end{bmatrix} \times \begin{bmatrix} U \\ V \\ W \end{bmatrix} + B_{bg} \begin{bmatrix} 0 \\ 0 \\ g \end{bmatrix}, \quad (2)$$

where $[U V W]^T$ is the velocity vector in the BFF, m is the mass of the SD-40, $[F_x F_y F_z]^T$ the total force vectors expressed in the BFF, and $[P Q R]^T$ is the vector of the attitude angular rates.



FIGURE 1: SD-40 UVA.

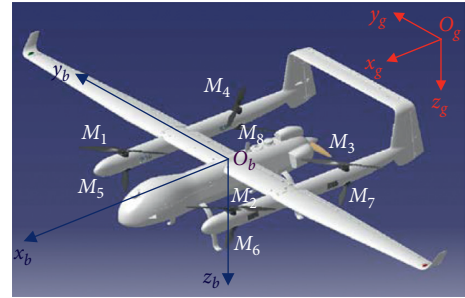


FIGURE 2: Reference coordinate systems for mathematical modeling.

$$\begin{bmatrix} F_x \\ F_y \\ F_z \end{bmatrix} = S_{ba} \begin{bmatrix} -D - F_D \\ Y \\ -L \end{bmatrix} + \begin{bmatrix} T \\ 0 \\ -F_T \end{bmatrix}, \quad (3)$$

$$S_{ba} = \begin{bmatrix} c\alpha c\beta & s\beta & s\alpha c\beta \\ -c\alpha s\beta & c\beta & -s\alpha s\beta \\ -s\alpha & 0 & c\alpha \end{bmatrix},$$

where S_{ba} is the transformation matrix from the airflow coordinate frame (AFF) to the BFF, α is the angle of attack, β is the sideslip angle, $[D Y L]^T$ is the total force vectors expressed in AFF, T is the horizontal thrust of the engine, F_T is the force in BFF produced by eight motors with rotors in multirotor mode, and F_D is the drag force in AFF produced by eight rotors in aircraft mode.

The components of the force vector and the drag force yields are expanded as

$$\begin{aligned} D &= qs * C_d, \\ L &= qs * C_l, \\ Y &= qs * C_y, \\ F_D &= K_d V_t^2, \end{aligned} \quad (4)$$

where V_t is the airspeed, K_d is the drag coefficient of rotors, qs is the dynamic pressure, and $[C_d C_l C_y]^T$ is the vector of the drag coefficient, the lift coefficient, and the side force coefficient of the fixed wing.

$[\text{Windu Windv Windw}]^T$ is the wind disturbance. The three-axis velocity in the BFF can be updated as

$$\begin{bmatrix} U_w \\ V_w \\ W_w \end{bmatrix} = \begin{bmatrix} U \\ V \\ W \end{bmatrix} - B_{bg} \begin{bmatrix} \text{Windu} \\ \text{Windv} \\ \text{Windw} \end{bmatrix}. \quad (5)$$

V_t , α , and β are equally updated as

$$\begin{aligned} V_t &= \sqrt{U_w^2 + V_w^2 + W_w^2}, \\ \alpha &= a \tan \frac{W_w}{U_w}, \\ \beta &= a \sin \frac{V_w}{V_m}. \end{aligned} \quad (6)$$

According to the relationship between the NED frame and the BFF, the kinematics equation of the mass center is obtained as

$$\begin{bmatrix} \dot{X} \\ \dot{Y} \\ -\dot{Z} \end{bmatrix} = B_{bg}^{-1} \begin{bmatrix} U_w \\ V_w \\ W_w \end{bmatrix}, \quad (7)$$

where $[X Y Z]^T$ represents the position vector of the UAV in the NED.

2.2. Angular Motion Equation. According to the moment of momentum theorem, the dynamic equation of the rotation of the mass center can be derived with respect to the inertia, the total moment of the fixed wing, and the angular rates as follows:

$$\begin{bmatrix} \dot{P} \\ \dot{Q} \\ \dot{R} \end{bmatrix} = -I^{-1} \left(\begin{bmatrix} P \\ Q \\ R \end{bmatrix} \times \left(I \begin{bmatrix} P \\ Q \\ R \end{bmatrix} \right) \right) + I^{-1} \begin{bmatrix} \bar{L} + M_{CL} \\ M + M_{CM} \\ N + M_{CN} \end{bmatrix} \quad (8)$$

where I is the inertia matrix of the rigid body and I^{-1} is the inverse matrix of I .

$$I = \begin{bmatrix} I_x & 0 & -I_{xz} \\ 0 & I_y & 0 \\ -I_{xz} & 0 & I_z \end{bmatrix}. \quad (9)$$

$[\bar{L} M N]^T$ is the total moment vectors of fix-wing expressed in the BFF, which can be expressed as

$$\begin{aligned} \bar{L} &= qs^* b^* C_r, \\ M &= qs^* c^* C_m, \\ N &= qs^* b^* C_n \end{aligned} \quad (10)$$

where b is the wingspan and c is the length of the average aerodynamic chord. $[C_r C_m C_n]^T$ is a vector representing the roll moment coefficient, the pitch moment coefficient, and the yaw moment coefficient of the fixed wing, respectively. $[M_{CL} M_{CM} M_{CN}]^T$ is the total moment vectors of the multirotor system expressed in the BFF as

$$\begin{aligned} M_{CL} &= -\frac{\sqrt{2}}{2} * l * (T_1 - T_2 - T_3 + T_4 + T_5 - T_6 - T_7 + T_8), \\ M_{CM} &= -\frac{\sqrt{2}}{2} * l * (-T_1 + T_2 + T_3 - T_4 - T_5 + T_6 + T_7 - T_8), \\ M_{CN} &= N_1 - N_2 + N_3 - N_4 - N_5 + N_6 - N_7 + N_8, \\ M_{Ni} &= K_n * N_i^2 (i = 1, 2, 3, 4, 5, 6, 7, 8), \end{aligned} \quad (11)$$

where l is the multirotor wheelbase, K_n is the rotational damping moment coefficient, N_i is the revolutions per minute (RPM) of the rotor, T_i is the force in the BFF produced by eight motors with rotors in multirotor mode, and M_{Ni} is the counter-torque of each rotor [11].

Based on the Euler Angle relation, the relationship between the attitude angle differential and the angular rate can be obtained as

$$\begin{bmatrix} \dot{\phi} \\ \dot{\theta} \\ \dot{\psi} \end{bmatrix} = \begin{bmatrix} 1 & s\phi \tan \theta & c\phi \tan \theta \\ 0 & c\phi & -s\phi \\ 0 & \frac{s\phi}{c\theta} & \frac{c\phi}{c\theta} \end{bmatrix} \begin{bmatrix} P \\ Q \\ R \end{bmatrix}. \quad (12)$$

The 6-DoF nonlinear dynamics equations can be expressed by equations (2), (7), (8), and (12).

3. Flight Controller Design

The flight controller of SD-40 consists of an inner loop for attitude control and an outer loop for trajectory control. The attitude angle error is used as the input signal for the attitude control. In contrast, the external loop controller uses the trajectory error and its rate of change as input signals.

3.1. Attitude Control. The attitude control design of the UAV can be carried out into extended state observer, pitch channel control, and roll channel control.

3.1.1. Extended State Observer. Equation 12 shows that the roll channel, pitch channel, and yaw channel are interrelated, which poses a challenge in the design of a conventional controller. However, Active Disturbance Rejection Control (ADRC) can solve this coupled problem. First, the interaction between different channels inside the system, together with the external disturbance caused by the environment, is treated as the total disturbance of the channel. Then, ESO is applied to each angle channel independently to estimate the total disturbance in real time. Finally, a total disturbance is used to dynamically compensate and linearize the UAV control system [12, 13].

Equation system (12) can be updated as

$$\begin{aligned}
\ddot{\phi} &= f_1(\phi, \dot{\phi}, \theta, \dot{\theta}, \psi, \dot{\psi}) + \omega_1 + b_1 \bar{L}, \\
\ddot{\theta} &= f_2(\phi, \dot{\phi}, \theta, \dot{\theta}, \psi, \dot{\psi}) + \omega_2 + b_2 M, \\
b_1 &= \bar{L}_{\delta a} = \frac{(qs * BL * Cl_{\delta a})}{I_x}, \\
b_2 &= M_{\delta e} = \frac{(qs * BA * Cm_{\delta e})}{I_y},
\end{aligned} \tag{13}$$

where $f_i(\phi, \dot{\phi}, \theta, \dot{\theta}, \psi, \dot{\psi})$ is a modeled nonlinear dynamic system and ω_i is the unmodeled dynamic system including external interference, modeling error, and other factors [14, 15]. b_1 is the control input gain from the deflection angle of the aileron angle to roll angular acceleration and b_2 is the control input gain from the deflection angle of the elevator angle to pitch angular acceleration. $Cl_{\delta a}$ is the coefficient of the rolling moment, while $Cm_{\delta e}$ is the coefficient of pitching moment [16].

The second-order system can be expanded into a third-order system as

$$\begin{cases} \dot{x}_1 = x_2 \\ \dot{x}_2 = x_3 + bu \\ \dot{x}_3 = \dot{f}(x_1, x_2, \omega(t)) \\ y = x_1 \end{cases} \tag{14}$$

The linear extended state observer can be unfolded into the following equation:

$$\begin{cases} \dot{e}_1 = z_1 - y \\ \dot{z}_1 = z_2 - \beta_1 e_1 \\ \dot{z}_2 = z_3 - \beta_2 e_1 + bu \\ \dot{z}_3 = -\beta_3 e_1 \end{cases} \tag{15}$$

where $[z_1 \ z_2 \ z_3]^T$ is the output of the ESO, tracking the state variables of the original system $[x_1 \ x_2 \ x_3]^T$ [17, 18]. All poles of the ESO characteristic equation can be located at $-\omega_0$ (on the left half-plane of the complex plane) as per the following equation:

$$\begin{cases} \beta_1 = 3\omega_0 \\ \beta_2 = 3\omega_0^2, \\ \beta_3 = \omega_0^3 \end{cases} \tag{16}$$

where ω_0 is denoted as the bandwidth of the ESO.

The control signal can be expressed as

$$u = \frac{u_0 - \beta \cdot z_3}{b}, \tag{17}$$

where u_0 is the output of PD controller and β is the compensation coefficient of total disturbance [19].

3.1.2. Pitch Channel Control Loop. The pitch channel controller needs the signals from the angular rate sensor and attitude angle sensor for feedback, increasing the damping of the UAV system and improving the dynamic response of the open-loop system.

The longitudinal control system structure of the SD-40 is shown in Figure 3, which is composed of the pitch attitude control loop based on ESO and the height control loop based on the total energy control system (TECS).

Pitch angle control based on ESO is an adaptive PD controller [20] with double loop feedback of the pitch angular rate Q and pitch angle θ . According to equations (14) and (15), x_1 is θ and x_2 is Q , z_1 is the estimation of the θ signal, z_2 is the estimation of the Q signal, and z_3 is the estimation of the uncertain parts.

The system block diagram is shown in Figure 4, and the mathematical expression is shown in the following equation: where δ_e is the deflection angle of the elevator, θ_g is the pitch angle command, and Q is the pitch angular rate feedback value.

$$\delta_e = K_e^\theta(\theta - \theta_g) + K_e^Q Q - \frac{\beta}{b_2} z_3, \tag{18}$$

b_2 can be calculated with equation (13) as $b_2 = (qs * BA * Cm_{\delta e})/I_y = -0.4 \sim -0.45$.

The bandwidth ω_0 of the pitch channel ESO can adjust the tracking speed of the observed state variable. According to reference [21], $\omega_0 \approx 5 - 10\bar{\omega}_c$, where $\bar{\omega}_c$ is the equivalent bandwidth of the transient profile, which is used with a settling time of 1 s, $\bar{\omega}_c = 4$, so $\omega_0 \approx 20 - 40$.

A higher bandwidth corresponds to a better command of the disturbance rejection and sensitivity to parameter variations. However, achievable bandwidth is limited by the hardware and software limitations, including sensor noises and the given sampling rate of the actual flight control hardware system. The comparison of responses at different values ω_0 is shown in Figure 5.

Considering the simulation results and the actual limitations, ω_0 is selected as 25. The system response versus the step input is shown in Figure 6. The pitch angle can quickly track instructions and can be well estimated by z1. Figure 7 is the observed value of total disturbance of pitch control loop. The disturbance can be suppressed in a short period.

3.1.3. Roll Channel Control Loop. Similar to the pitch channel, the roll angle control is also developed based on the PD philosophy. The control block diagram of the roll channel is shown in Figure 8.

The control law is as follows:

$$\delta_a = K_a^\phi(\phi - \phi_g) + K_a^P P - \frac{\beta}{b_1} z_3, \tag{19}$$

where δ_a is the deflection angle of the aileron, ϕ_g is the roll angle command, and P is the roll angular rate feedback value, improving the short period damping.

b_1 can be calculated with equation (13) [12]: $b_1 = (qs * BL * Cl_{\delta a})/I_x = -0.4 \sim -0.49$.

The comparison of response at different values ω_0 is shown in Figure 9. As done previously for the pitch channel, the bandwidth of the roll channel ESO ω_0 is selected as equal to 25.

The roll angle step response is shown in Figure 10, which has good tracking and observation effects. Figure 11 is the

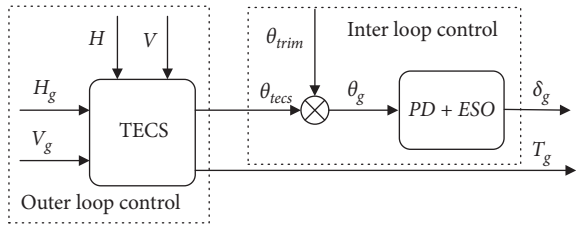


FIGURE 3: The longitudinal control structure of the SD-40.

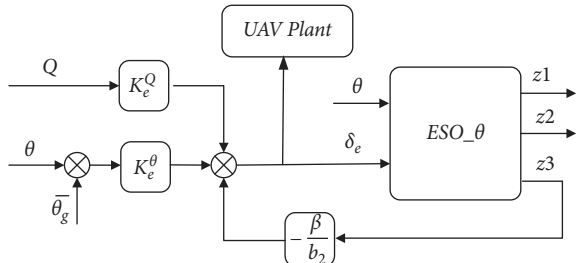


FIGURE 4: Pitch angle controller block diagram of the SD-40.

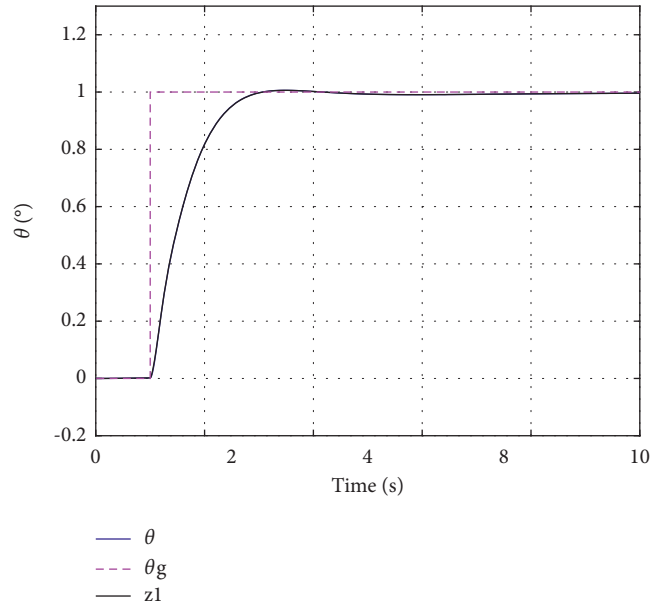


FIGURE 6: Pitch angle step response ($\omega_0 = 25$).

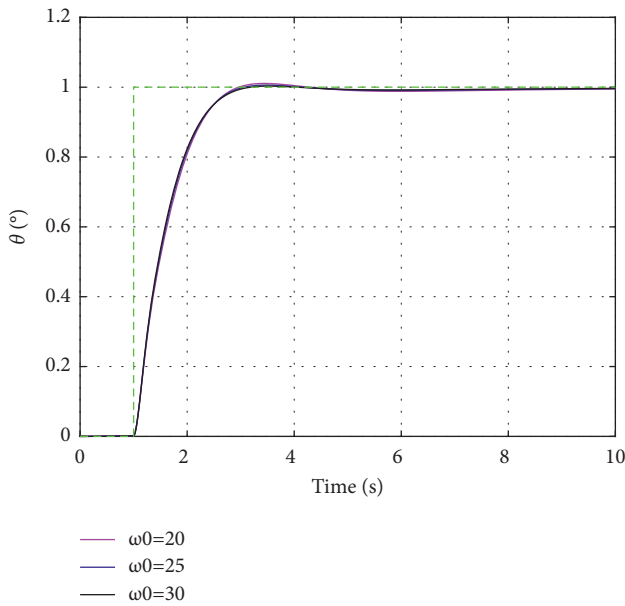


FIGURE 5: Comparison of response at different values of ω_0 .

observed value of the total disturbance of the Roll control loop. The disturbance can be suppressed in a short period and the system can quickly reach a stable state.

The pitch angle and the roll angle controller based on ESO for the inner loop can respond to and track the command effectively and achieve high precision positioning of the UAV.

3.2. Trajectory Control

3.2.1. Longitudinal Height Control. The control of the longitudinal height of the SD-40 is performed through an outer loop based on the TECS. The throttle controls the change of total energy, and the deflection angle of the

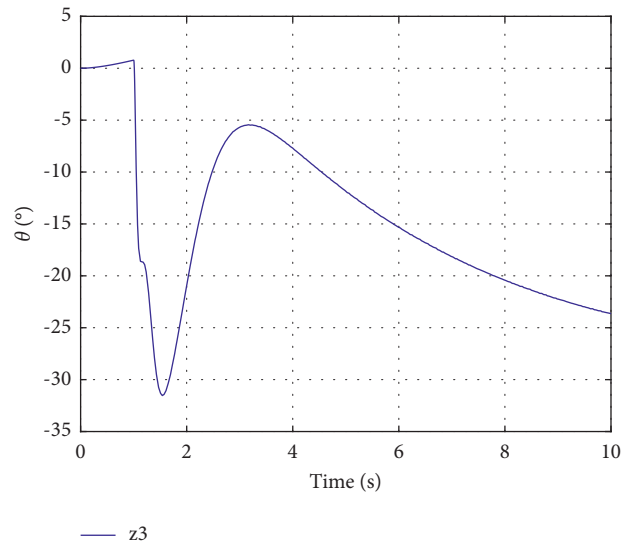


FIGURE 7: Total disturbance observation.

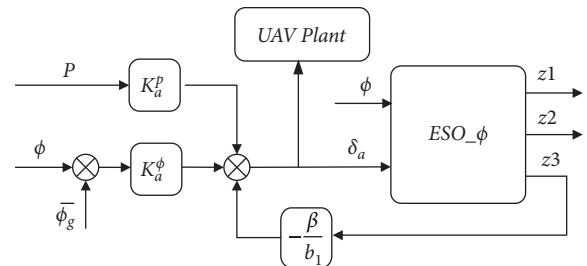
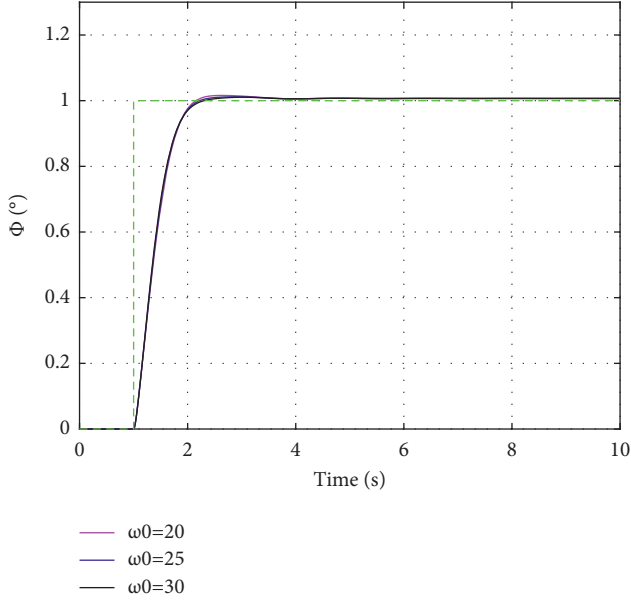
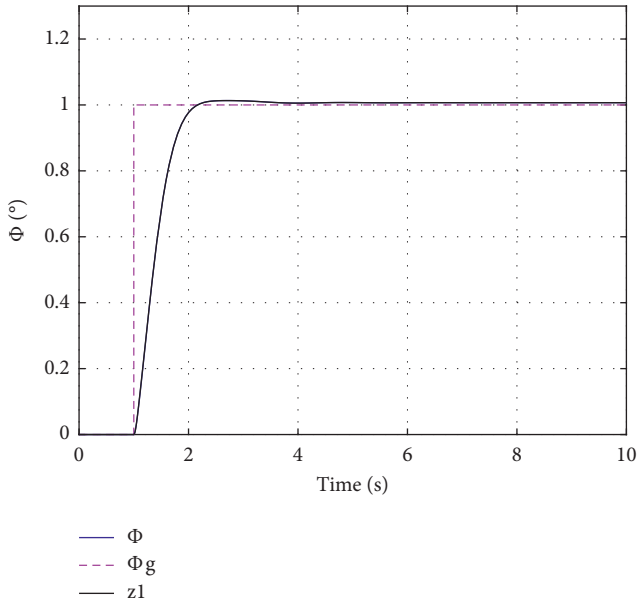


FIGURE 8: Roll angle controller block diagram of the SD-40.

elevator distributes the kinetic energy and the gravitational potential energy. As such, the speed control and the height control are decoupled [22–24].

FIGURE 9: Comparison of response at different values of ω_0 .FIGURE 10: Roll angle step response ($\omega_0 = 25$).

The error between the kinetic energy and the specified kinetic energy at a certain moment can be expressed as

$$K_{\text{err}} = K_{\text{ref}} - K = \frac{1}{2}m(V_t^c)^2 - \frac{1}{2}mV_t^2, \quad (20)$$

where K is the kinetic energy of the particle, V_t is the true airspeed, and V_t^c is the specified airspeed.

The error between the gravitational potential energy and the specified gravitational potential energy at a certain moment can be expressed as

$$U_{\text{err}} = mg(h^c - h). \quad (21)$$

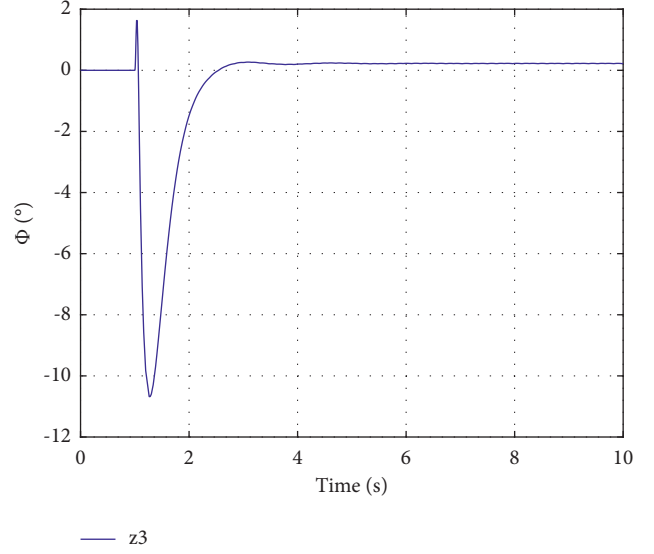


FIGURE 11: Total disturbance observation.

The error of the total energy E and the allocation L of the total energy error can be expressed as per the following equation:

$$\begin{aligned} E &= U_{\text{err}} + K_{\text{err}}, \\ L &= U_{\text{err}} - K_{\text{err}}. \end{aligned} \quad (22)$$

The allocation is updated as

$$L = (2 - k)U_{\text{err}} - kK_{\text{err}}. \quad (23)$$

By differentiating both sides of the total energy equation and assuming that the trajectory angle γ is small, the rate of change of the total energy is approximately expressed as

$$\dot{E} = \frac{\dot{E}_T}{mgV_t} = \frac{\dot{V}_t}{g} + \frac{\dot{h}}{V_t} = \frac{\dot{V}_t}{g} + \sin \gamma \approx \frac{\dot{V}_t}{g} + \gamma. \quad (24)$$

Meanwhile, the longitudinal dynamics equation of the UAV is expressed as per the following equation:

$$T - D = m\dot{V}_t + mg \sin \gamma \approx mg \left(\frac{\dot{V}_t}{g} + \gamma \right), \quad (25)$$

where T and D are, respectively, the thrust and drag force of UAV during cruise state. In this balanced state, the thrust increment is expressed as

$$\Delta T = mg \left(\frac{\dot{V}_t}{g} + \gamma \right) = mg \frac{\Delta \dot{E}}{V}. \quad (26)$$

In summary, the change of total energy can be controlled by the change in thrust [25].

The deflection angle of the elevator changes the track angle, which leads to the mutual conversion between the potential energy and the kinetic energy [26, 27]. Assuming that there is no significant energy loss during the conversion process, the change rate of the energy distribution can be defined as

$$\dot{L} = \gamma - \frac{\dot{V}_t}{g}. \quad (27)$$

Total energy control block diagram is shown in Figures 12 and 13.

H_g is the height command, V_g is the speed command. E_{sp} is the specified energy, L_{sp} is the specified energy distribution, and \dot{E}_{sp} and \dot{L}_{sp} are the corresponding rates of change. E is the total energy, L is the energy distribution, and \dot{E} and \dot{L} are the corresponding rates of change, respectively. $K_{\dot{E} \rightarrow \Delta T}$ is the gain from energy change rate to throttle and $K_{\dot{L} \rightarrow \Delta \theta}$ is the gain from energy distribution change rate to pitch angle.

The complete TECS control law, as well as the total energy change control channel, can be expressed as

$$\begin{aligned} \dot{E}^c &= k_{pE} \times \left(E + d_{pE} \dot{E} + I_{pE} \int E dt \right), \\ T_g &= \dot{E}^c \times K_{\dot{E} \rightarrow \Delta T}, \end{aligned} \quad (28)$$

$$K_{\dot{E} \rightarrow \Delta T} = \frac{\delta_{T_{\max}} - \delta_{T_{\min}}}{\dot{E}_{\max} - \dot{E}_{\min}},$$

where \dot{E}_{\max} and \dot{E}_{\min} are the max and min rate of change of the total energy in climbing mode and descending mode, respectively. $\delta_{T_{\max}}$ and $\delta_{T_{\min}}$ are the max and min opening of the throttle, respectively.

The total energy distribution control channel could then be expressed as

$$\dot{L}^c = k_{eL} \times \left(L + d_{eL} \dot{L} + I_{eL} \int L dt \right). \quad (29)$$

Controlling \dot{L}^c with the elevator requires a pitch rate/pitch attitude feedback inner-loop control law, where θ_g will be transformed into elevator command to stabilize the UAV's short period mode. Thus, \dot{L}^c develops θ_g , serves as the input of the pitch attitude control loop, which acts indirectly on the elevator [28–30].

$$\frac{\dot{L}}{gV_t} = \frac{gh - V_t \dot{V}_t}{gV_t} \approx \gamma - \frac{\dot{V}_t}{g}, \quad (30)$$

$$\theta_g = \frac{\dot{L}^c}{K_{\dot{L} \rightarrow \Delta \theta}} = \frac{1}{gV_t} \times \dot{L}^c,$$

where gV_t is the conversion coefficient from the climb angle to energy distribution change rate, which adapts to a range of speed changes.

(1) Simulation Data.

Case 1: Given the initial state of UAV, height is 100 m and speed is 30 m/s. When $t=1$ s, change the height command to 105 m, then the height step response is shown in Figure 14.

From the graph in Figure 15, the height can track the command well with some overshoot. Meanwhile, the speed is basically unchanged, and the fluctuation is controlled within ± 0.5 m/s.

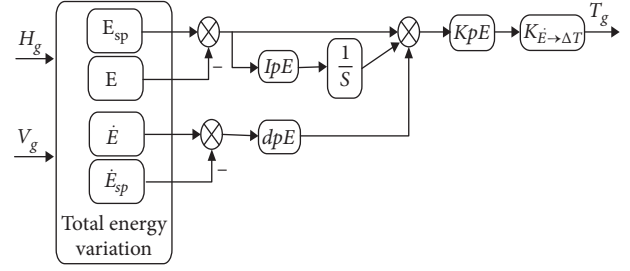


FIGURE 12: Total energy change control.

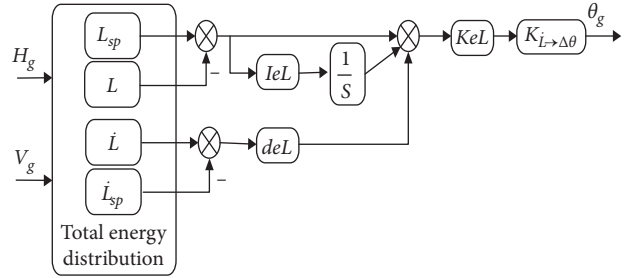


FIGURE 13: Total energy distribution control.

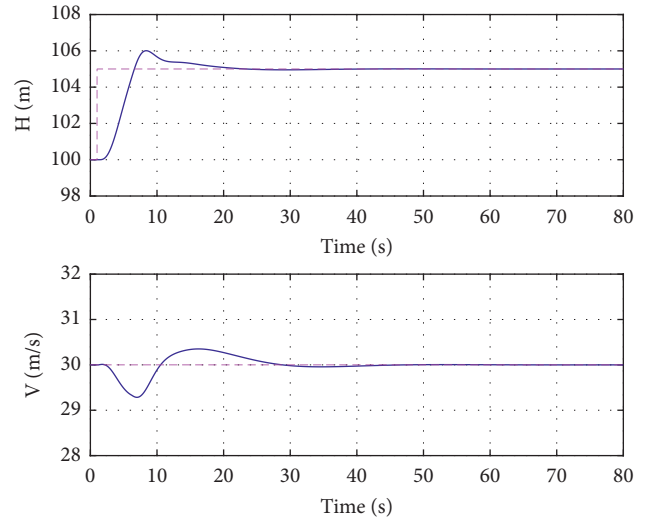


FIGURE 14: Height step response for a given speed.

As the height increases, the kinetic energy is converted into potential energy, which requires a larger throttle energy value. Even though the value of the elevator energy is negative, the UAV performs an upward movement. When the new balance is reached, the value comes back to zero.

Case 2: Given the initial state of UAV, the height is set equal to 100 m and the speed to 30 m/s. When time $t=1$ s, change the speed command to 32 m/s, then speed step response is shown in Figure 16.

As shown in Figure 17, the speed can track the command well, with some likely overshoot. Meanwhile, the height is unchanged, and the fluctuation is controlled within a distance of ± 0.3 m.

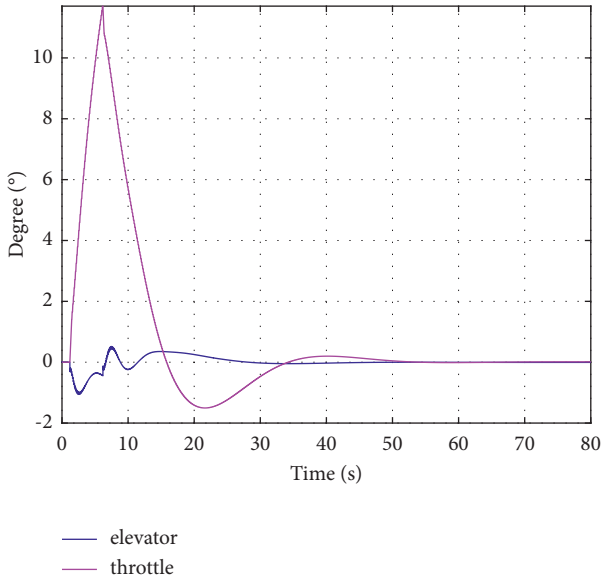


FIGURE 15: Elevator and throttle command for height step response simulation.

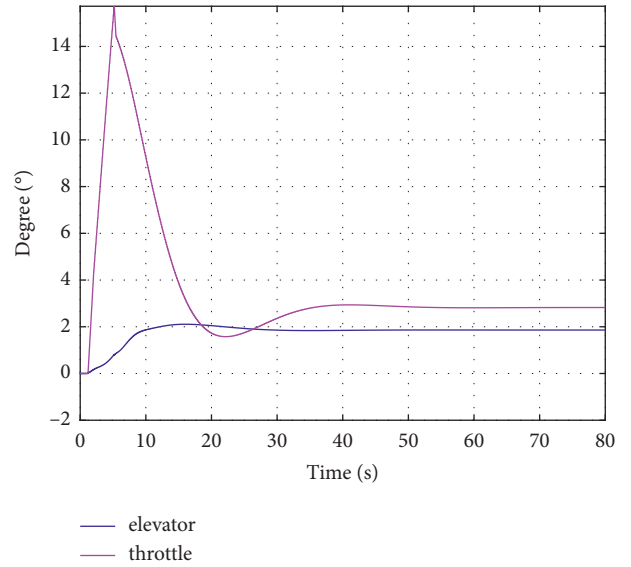


FIGURE 17: Elevator and throttle command for speed step response simulation.

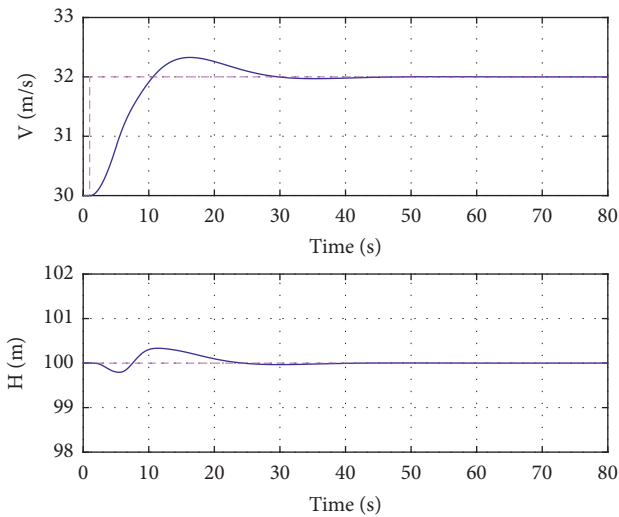


FIGURE 16: Speed step response for a given height.

As the speed increases, the kinetic energy increases, requiring a larger throttle energy value. The elevator's value is positive to accelerate the UAV, and when the new balance is reached, the value keeps at 2° , as shown in Figure 15.

(2) *Flight Data.* The UAV's initial height is 1470 m and speed is 30 m/s. When $t = 1$ s, the change in height command and the speed command are $H_g = 1170$ m, $V_g = 32$ m/s, then step response is shown in Figure 18.

As shown in Figures 18 and 19, the altitude drops slowly with the command, and the airspeed is kept at 35 ± 1 m. Since the UAV flies at a constant speed, the throttle remains unchanged to keep the kinetic energy intact. The elevator command is essentially negative,

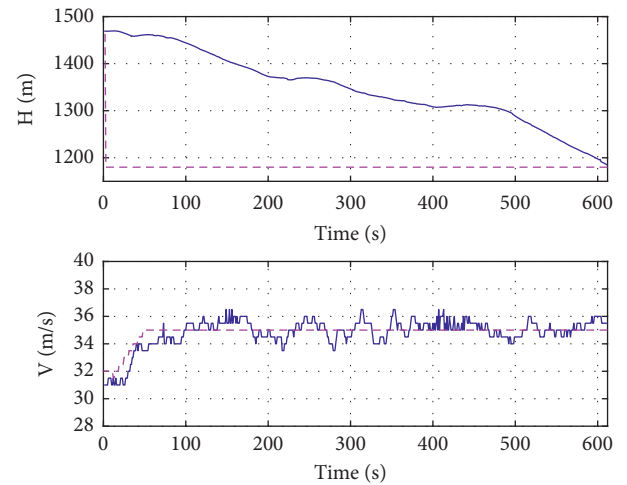


FIGURE 18: Height and speed response of step input test.

raising the UAV's head and increasing the drag as the gravitational potential energy slowly converts to internal power to work done against friction.

3.2.2. *Lateral Trajectory Control.* The lateral trajectory consists of the roll and yaw motions, which have been considered in the controller design. L1 guidance is well known for its simplicity in tracking circular and linear motions typical for fixed-wing vehicles' flight plans. However, the law has a specific ADRC effect in curve tracking, which is introduced for circular trajectory tracking, as shown in Figure 20. Under this section, the controller design is presented along with simulated results and the profile of the flight data.

When the UAV moves laterally, the roll angle is not zero, resulting in lift tilt and lateral acceleration as expressed by the following equation:

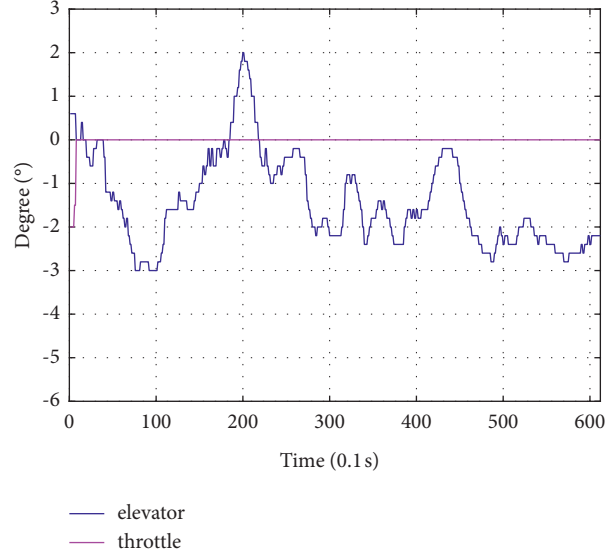


FIGURE 19: Elevator and throttle command for height and speed step response test.

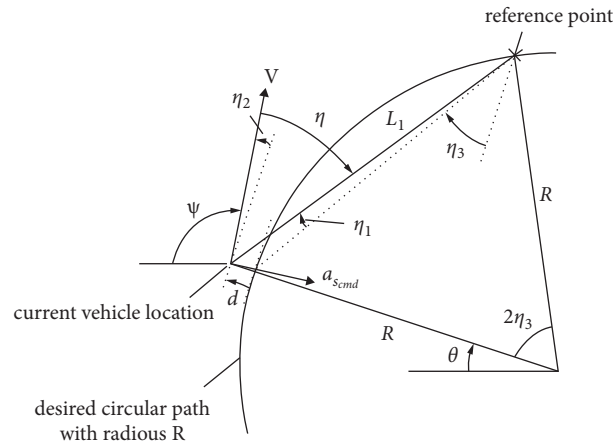


FIGURE 20: Principle diagram of circular trajectory tracking.

$$a_{s_{cmd}} = \frac{V^2}{R} = \frac{V^2}{L_1/2 \sin \eta} = \frac{2V^2}{L_1} \sin \eta, \quad (31)$$

$$\sin \eta = \sin(\eta_1 + \eta_2 + \eta_3) \approx (\eta_1 \cos \eta_3 + \eta_2 \cos \eta_3 + \sin \eta_3), \quad (32)$$

$$a_{s_{cmd}} = \frac{2V^2 c^2}{L_1^2} d + \frac{2Vc}{L_1} \dot{d} + \frac{V^2}{R}. \quad (33)$$

where $a_{s_{cmd}}$ is the acceleration command, L_1 is the distance between the reference point and the UAV, which is a fixed value, and η is the angle from the ground speed to the line L_1 . Meanwhile, the velocity direction change due to the acceleration can be expressed as

$$\Delta \psi = \frac{a_s}{V} \Delta t, \quad (34)$$

$$a_{s_{cmd}} \approx V \dot{\psi} = V(\dot{\theta} - \dot{\eta}_2) = \frac{V^2}{R} - \ddot{d}. \quad (35)$$

Based on the equations (33) and (35), we can get

$$\ddot{d} + 2\zeta\omega_n \dot{d} + \omega_n^2 d \approx 0, \quad (36)$$

where $\zeta = \sqrt{2}/2$, $\omega_n = \sqrt{2}Vc/L_1$, and $c = \sqrt{1 - (L_1/2R)^2}$.

The undamped natural frequency is determined by V and L_1 :

$$L_1 = \frac{VTc}{\sqrt{2\pi}}. \quad (37)$$

The period and damping ratio are taken as the control parameters of the control law and L_1 as a variable proportional to the flight speed [31].

The lateral acceleration is provided by the lateral component of the lift force during the rolling of the UAV as

$$a_{s_{cmd}} = g \cdot \tan \phi. \quad (38)$$

The roll angle command can be expressed as

$$\phi_{cmd} = \tan^{-1} \left(\frac{a_{s_{cmd}}}{g} \right). \quad (39)$$

L_1 control structure diagram shown in Figure 21 is designed based on the following equation:

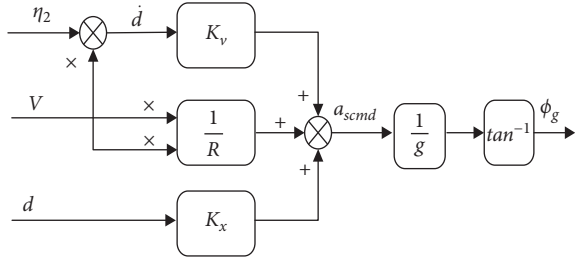


FIGURE 21: L1 block diagram. Here, η_2 is the yaw angle from the target heading to the speed heading and d is the lateral deviation distance, that is, the distance between the UAV real-time position and the center of the tracked circular arc minus the radius of the tracked circle.

$$a_{s_{cmd}} = K_x d + K_v \dot{d} + \frac{V^2}{R}, \quad (40)$$

where $K_x = \omega_n^2 = 4\pi^2/T^2$, $K_v = 2\zeta\omega_n = 4\pi\zeta/T$, and R is the radius of the tracked circular arc.

(1) *Tracking Simulation.* A nonlinear model in Matlab/Simulink environment has been adopted to carry out the circular trajectory tracking simulation. Given the L1 guidance law parameters period damping $\zeta = 0.707$, the simulation data at different periods T , as shown in Figures 22 and 23.

It can be seen from Figures 22 and 23, with the decrease of the period T , the response time of the lateral deviation distance and yaw angle error, that the error rapidly narrows, and the steady-state error of the lateral deviation distance decreases accordingly. In contrast, the yaw angle error fluctuates significantly. Reducing the period T will make the system response faster and the control effect more practical. A smaller value of the period T will lead to more aggressive navigation or sharper corners, while a more considerable value will lead to gentler navigation. Meanwhile, considering the safety and control rapidity in the UAV operation, the value of T is selected here as equal to 20 s.

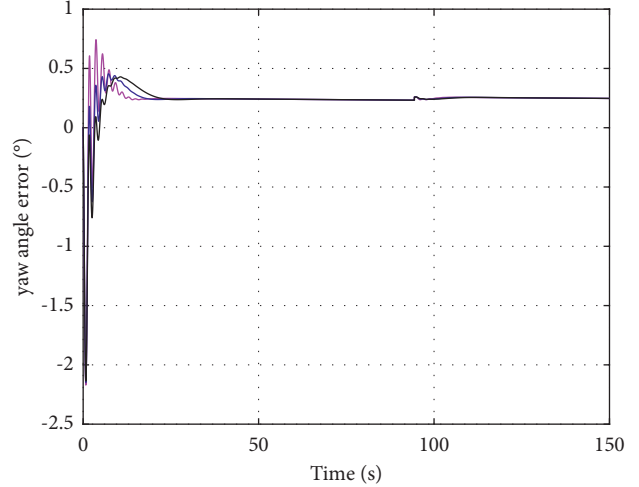
Setting $\zeta = 0.707$ and $T = 20$ s simulation initial state $V_t = 30$ m/s, constant wind disturbances in the north direction and east direction are added $U = 5$ m/s and $V = 5$ m/s. Track a circular trajectory with a radius of 600 m ($R = 600$ m).

The velocity profile with wind disturbances is shown in Figure 24. The tracking effect is shown in Figures 25 and 26.

As shown in Figure 24, the crosswind with 5 m/s has been simulated.

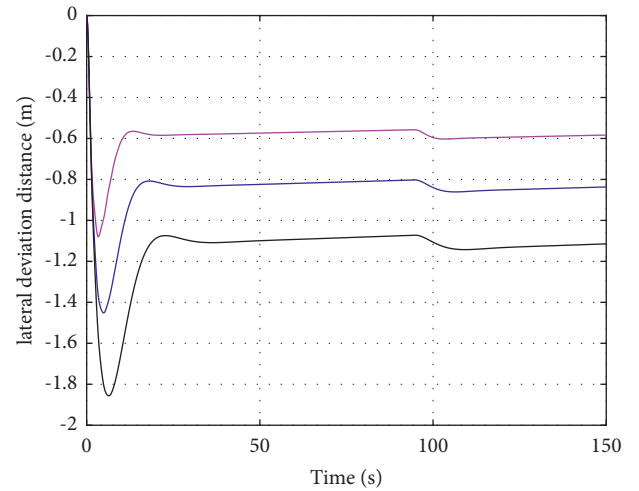
Simulation results show that the lateral deviation distance can be controlled at ± 3 m precision, while the yaw angle error is $\pm 0.5^\circ$, which offers a good effect of the circular trajectory tracking.

Figures 27 and 28 show the simulation of circular track tracking in a windless environment. Setting $\zeta = 0.707$ and $T = 20$ s, initial state $V_t = 30$ m/s, track a circular trajectory with a radius of 600 m. The lateral deviation distance is controlled at -1 m, while the yaw angle error is 0.4° , which shows a good effect of circular trajectory tracking without wind disturbance.



— zeta=0.707,T=15
— zeta=0.707,T=20
— zeta=0.707,T=25

FIGURE 22: Yaw angle error at different periods.



— zeta=0.707,T=15
— zeta=0.707,T=20
— zeta=0.707,T=25

FIGURE 23: Lateral deviation distance at different periods.

(2) *Flight Data Profile.* The UAV tracks a circle with a radius of $R = 450$ m, and the center of the circle is located at longitude: 100.23° , latitude: 41.01° , the tracking effect is shown in Figure 29.

Under the constant wind disturbance of 7 m/s (direction 184), airspeed values vary from 30 m/s to 37 m/s, as shown in Figure 30.

The value of the lateral deviation distance is between 0 and -4 m (shown in Figure 31), which means the UAV keeps on the left of the desired circular path. The yaw angle error can be controlled at $\pm 3^\circ$ (shown in Figure 32), larger than the value obtained by simulation. It is affected by crosswind to some extent, but the overall tracking effect is relatively good.

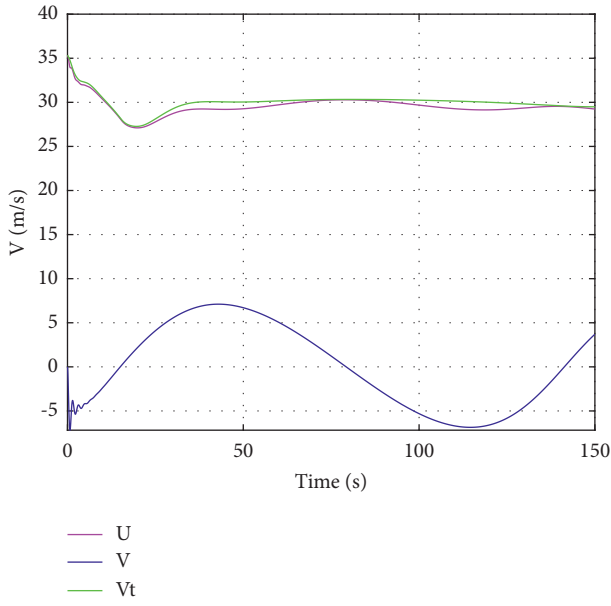


FIGURE 24: Velocity profile with wind disturbances.

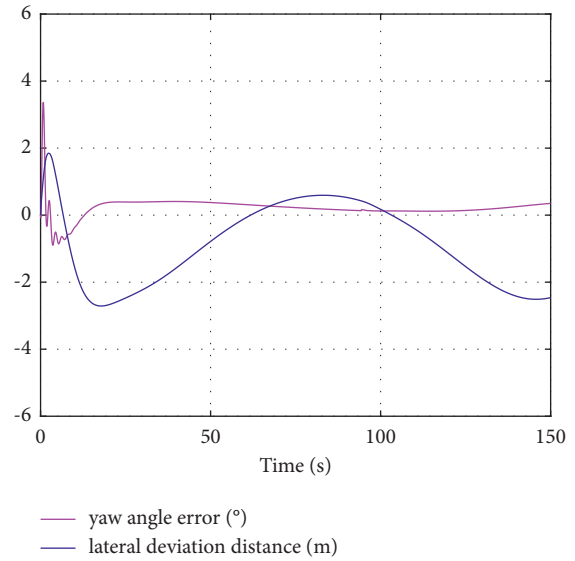


FIGURE 26: Angle error and deviation distance under wind disturbance.

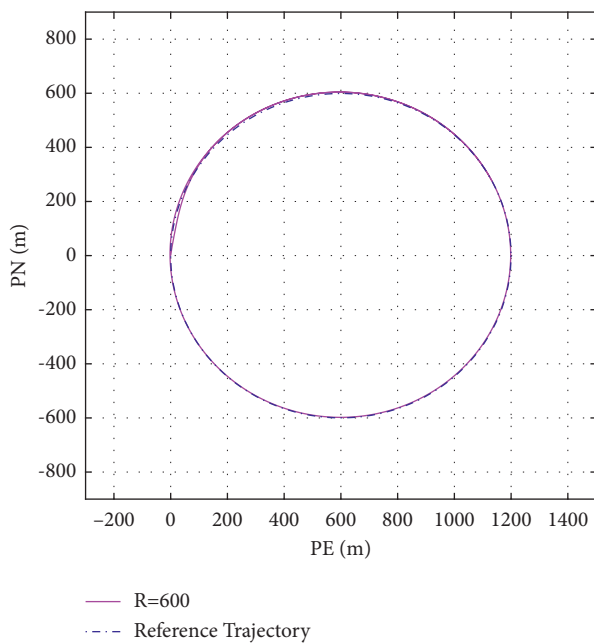


FIGURE 25: Circular trajectory tracking under wind disturbance.

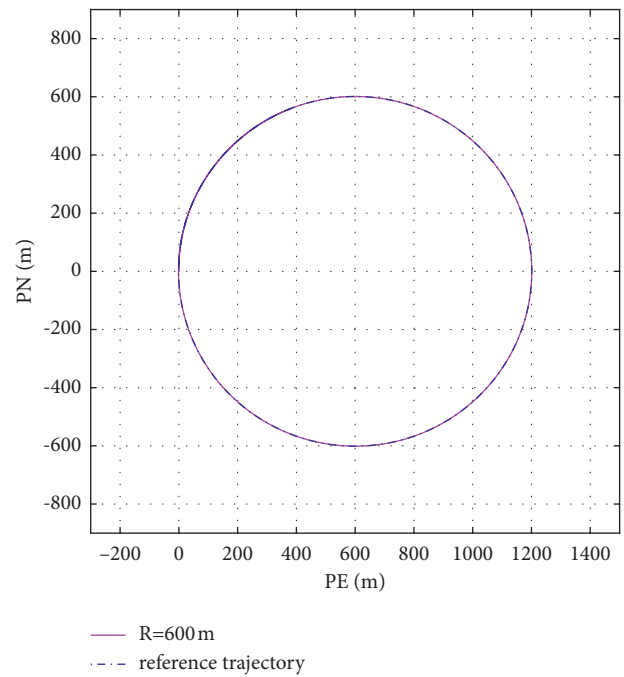


FIGURE 27: Circular trajectory tracking.

For the SD-40, $L1$ guidance law enhances the circular trajectory tracking control ability and the ability to resist wind, using the insensitive period and damping parameters. This law can calculate different expected accelerations according to different speeds V and offers good adaptability to different usual flight speeds and interference of ambient wind in actual flight. Furthermore, $L1$ guidance law can track irregular curves and be extended to more advanced trajectory algorithms, including obstacle avoidance flight and formation flight.

3.3. *Emergency Control.* Due to the limitations of cost, payload weight, and payload size, most small civil UAVs

adopt the Attitude and Heading Reference System (AHRS) based on a microelectromechanical system (MEMS). AHRS cannot provide accurate and stable attitude data under the following conditions:

- (a) When the UAV is under large continuous overload
- (b) When the algorithm of attitude solving unit has some problem
- (c) When the calculated data of attitude solving unit are divergent
- (d) When the magnetometer is interfered with and provides false heading information

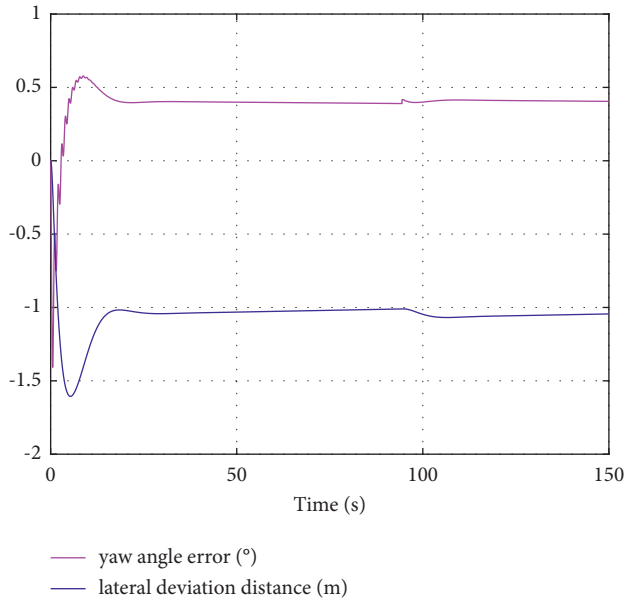


FIGURE 28: Angle error and deviation distance.

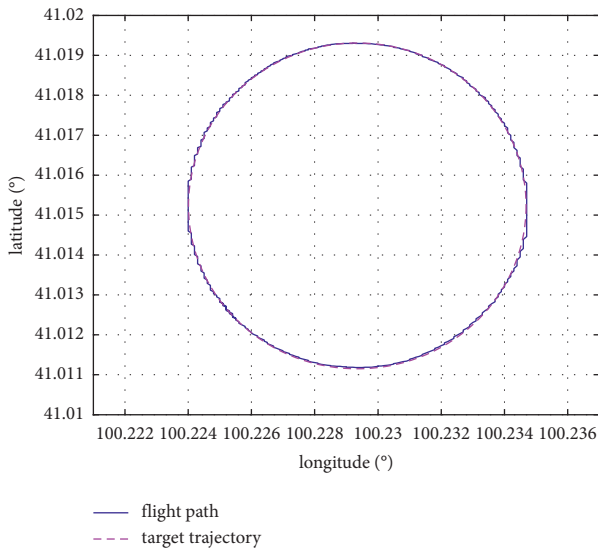


FIGURE 29: Circular trajectory tracking for the flight test.

In all these four cases, a triaxial attitude angle information and geomagnetic course information are invalid. The control method is the same in all the four cases because it is designed to deal with this information invalidation. In this study, we intend to propose a general solution for this information invalidation. The traditional attitude angle control technology is no longer applicable, and the control law based on the angular rate is needed [32]. Sound signals for flight controllers are UAV's position information, track angle, and ground speed provided by GPS; airspeed, atmospheric altitude, and rate of change information supplied by atmospheric pressure sensor; triaxial acceleration and triaxial angular rate provided by AHRS. The emergency controller design looked at the longitudinal and lateral control of the UAV.

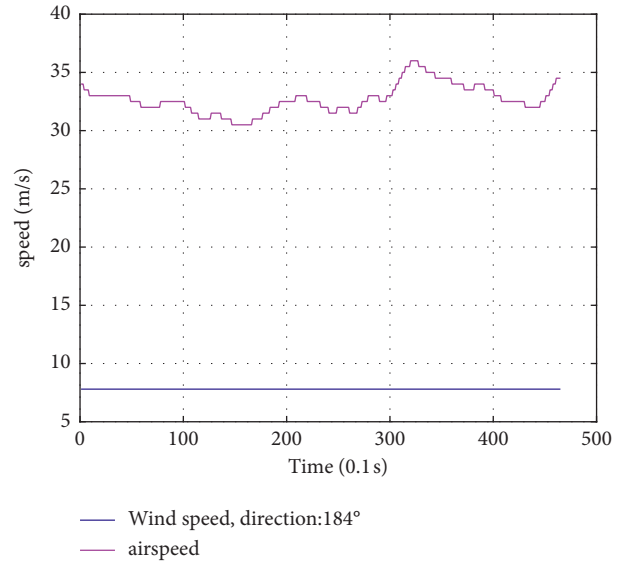


FIGURE 30: Wind speed and airspeed for the flight test.

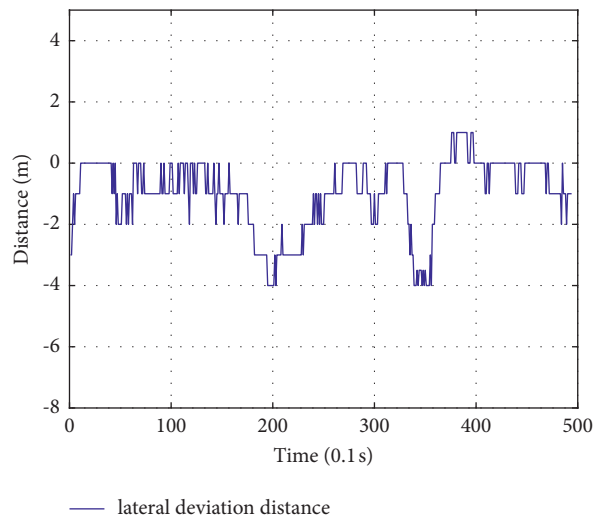


FIGURE 31: Lateral deviation distance for the flight test.

3.3.1. Longitudinal Control Law. TECS can decouple the height control loop and the speed control loop of the longitudinal motion, while the latter has no attitude control loop but directly controls the engine's thrust. Therefore, the speed control loop of TECS law can continue to be used. On the other hand, the height control law needs to be updated.

The pitch control loop is used to stabilize the pitch attitude and serves as the inner loop of the height controller. Since there is no pitch angle feedback information, height control based on angular pitch rate is proposed. The reduced-order ESO model is used for the inner-loop angular rate control. The longitudinal controller structure is shown in Figure 33:

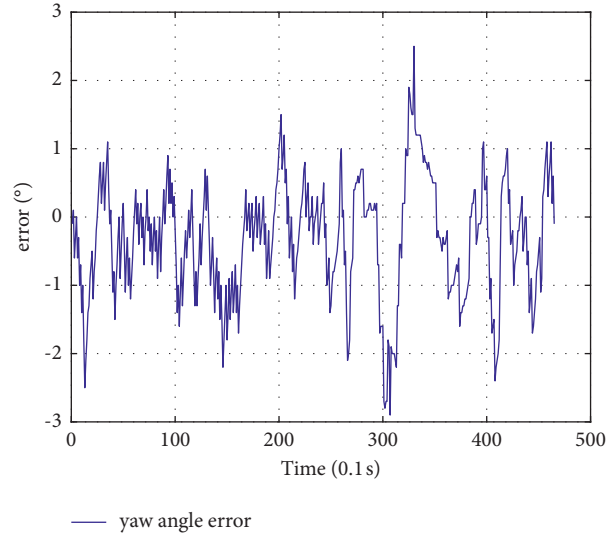


FIGURE 32: Yaw angle error for the flight test.

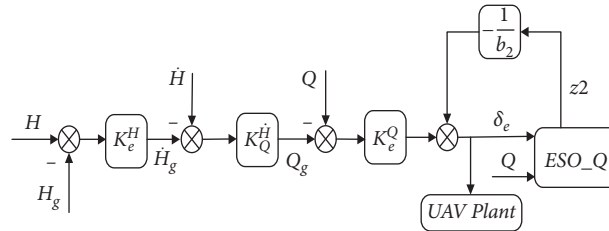


FIGURE 33: Longitudinal controller structure of the SD-40.

$$\begin{aligned} \delta_e &= K_e^{Iq} \int (Q - Q_g) dt + K_e^Q Q, \\ Q_g &= K_Q^{\dot{H}} [\dot{H} - K_e^H (H - H_g)], \end{aligned} \quad (41)$$

where Q_g is the given pitch angular rate and \dot{H} is the vertical speed of the UAV.

The outer loop height control output serves as the input to the inner-loop attitude angular rate control loop, which ultimately acts on the elevator.

The UAV's initial state of the simulation: height is 100 m and speed is 30 m/s. When $t = 1$ s, the height command is changed to $H_g = 105$ m.

The vertical speed is obtained according to a specific proportional relation of the error between height and its command. When $t = 1$ s, there is a sudden change in the vertical speed command due to the considerable height error, but the vertical speed can track the command well and respond quickly (Figure 34).

As shown in Figure 35, the value of the elevator energy is negative to enable the UAV to move upward, and when the new balance is reached, the energy value comes back to zero.

The height step command is tracked without overshoot, and the response is quick (see Figure 36). In addition, by putting an integral part into the pitch angular rate control loop, the influence of external interference on the system can be easily neglected.

3.3.2. Lateral Control Law. The structure of lateral roll angular rate control is similar to that of longitudinal pitch angular rate control. The turning control of UAV is realized by controlling the yaw angular rate.

The outer track control loop outputs the roll angular rate command, which will be used as input to the inner loop of the roll to control the aileron deflection angle (see Figure 37 and equations (42) and (43)):

$$\delta_a = K_a^{I_P} \int (P - P_g) dt + K_a^P P, \quad (42)$$

$$P_g = K_a^R [R - K_a^{P_{psi}} (Psi - Psi_g)], \quad (43)$$

where P_g is the roll angular rate command, R is the feedback of yaw angular rate, and Psi is the feedback of the track azimuth, which can be provided by GPS.

As shown in Figure 38, the track azimuth tracked the command well without overshoot.

The yaw angular rate command needs to be optimized into a smooth transient profile (Figure 39), as the desired trajectory to allow the output to follow more closely.

It can be seen from Figures 40 and 41 that the proportion term plays a significant role in the roll angular rate control loop. In addition, simulation data show that when AHRS fails, the emergency control law without Euler Angle feedback can track the control instructions well and enable the UAV to make a crash landing or continue to perform the mission, improving flight safety.

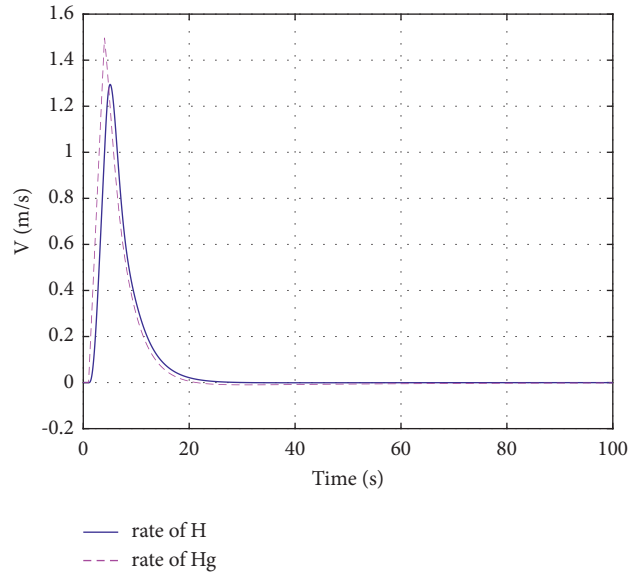


FIGURE 34: Vertical speed response to height changes.

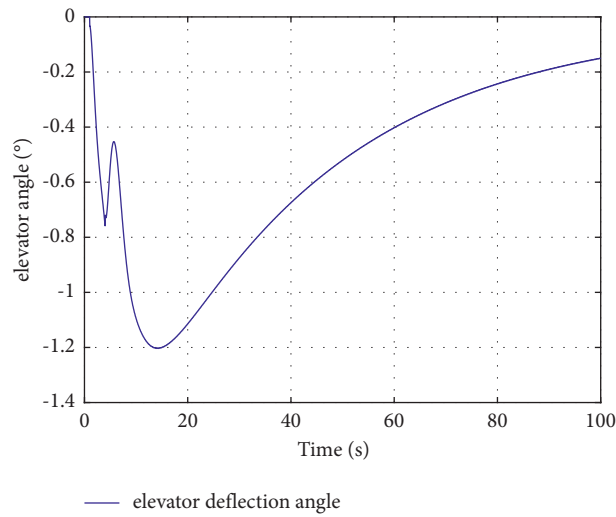


FIGURE 35: Elevator deflection angle response to height changes.

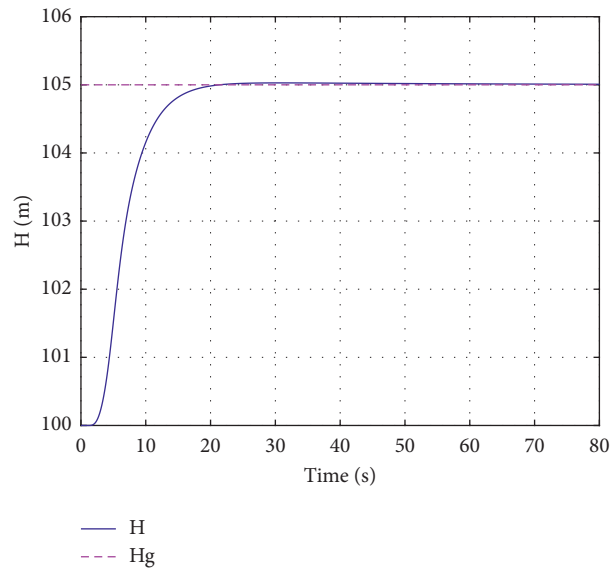


FIGURE 36: Height step response of the SD-40.

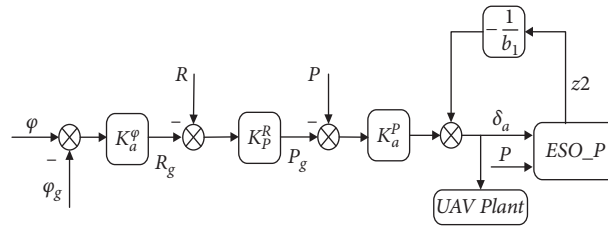


FIGURE 37: Lateral controller structure of the SD-40.

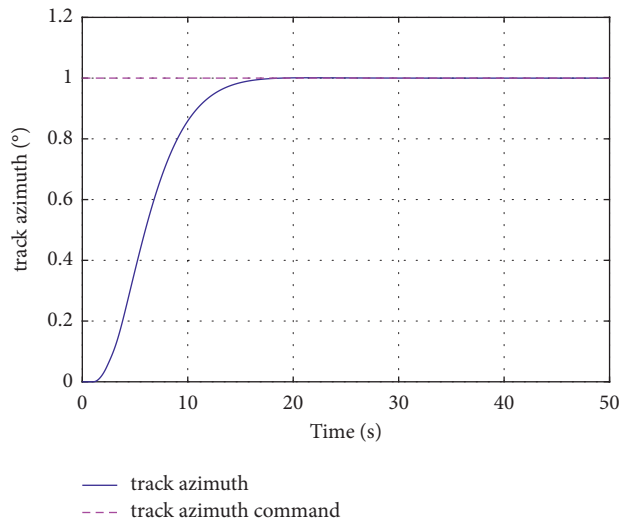


FIGURE 38: Track azimuth step response to command tracking.

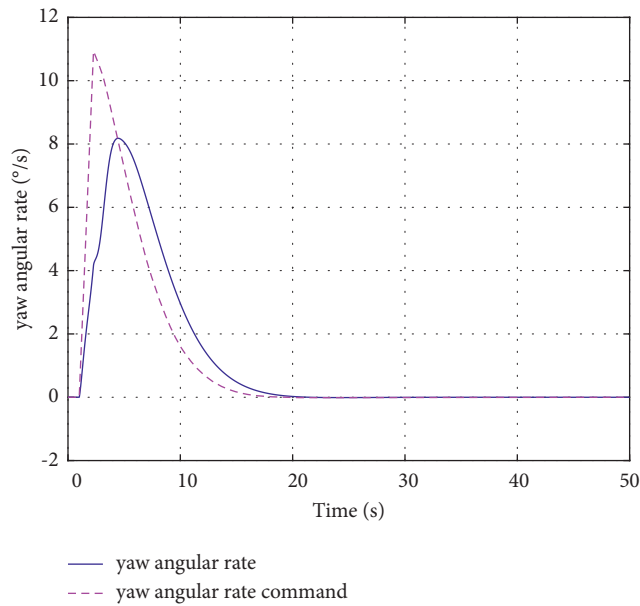


FIGURE 39: Yaw angular rate response to command tracking.

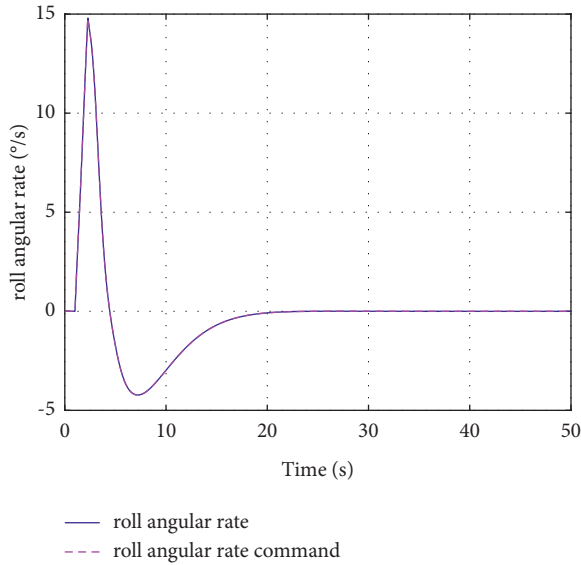


FIGURE 40: Roll angular rate response to command tracking.

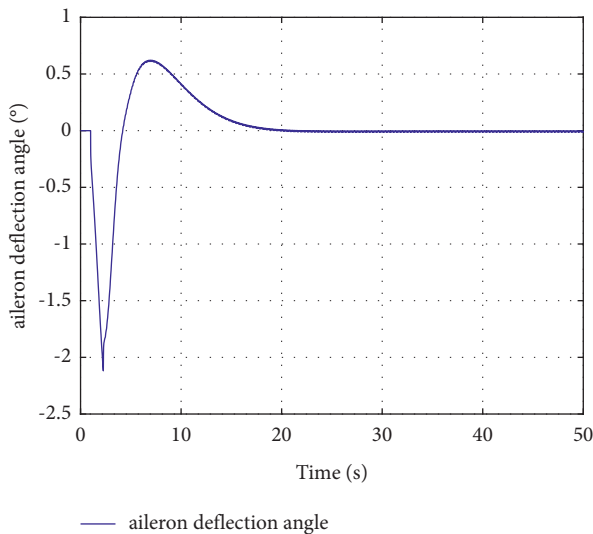


FIGURE 41: Aileron deflection angle response to command tracking.

4. Conclusion

In this paper, an aircraft flight mode controller for fixed-wing UAVs with VTOL functions has been successfully designed where the SD-40 served as a reference. The controller design process starts by establishing linear and angular perspective, then proceeds with the design of flight controller. Controllers for the attitude, the trajectory, and emergency condition flight have been developed and tested through simulations in the MATLAB and SIMULINK environment. First, pitch and roll channel controllers have been designed for attitude control. Next, the total energy control system, the extended state observer, and L1 guidance have been proposed for longitudinal and lateral control of fixed-wing UAV SD-40. An attitude controller based on the ESO has been proved to be effective from its anti-interference ability. The TECS is again proposed for the

longitudinal height channel, while the L1 guidance law is used for lateral trajectory tracking. The controller and simulation data show that the controller effectively responds to the command and decouples the speed control and the track control while improving the curve tracking ability and the wind resistance ability. Finally, a controller based on angular rate has been proposed in emergency and simulation data to verify its effectiveness.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Authors' Contributions

Deng Zhao performed the study, collected the data, and wrote the manuscript. Professor Wu conceived the idea of the study and designed the study. Professor You conceived the idea of the study and provided equipment and instruments for the research. All authors contributed to analyze the data and revisions.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 61673327).

References

- [1] G. Wu, B. Zhou, and L. Yang, "Review and prospect of domestic and foreign civil UAV industry development," *Economic Research Guide*, no. 12, pp. 160–162, 2016.
- [2] C. Roman, M. Lemanowicz, Z. Gorol, and T. Kudala, "Construction prototyping, flight dynamics modeling, and aerodynamic analysis of hybrid VTOL unmanned aircraft," *Journal of Advanced Transportation*, vol. 2018, Article ID 7040531, 15 pages, 2018.
- [3] K. Kenta and M. Masafumi, "Development of Tailsitter VTOL using multi rotor system," in *Proceedings of the JSME annual Conference on Robotics and Mechatronics, 2017*, Tokyo, Japan, November 2017.
- [4] K. Dalamagkidis, "Aviation history and unmanned flight," *Handbook of Unmanned Aerial Vehicles*, Springer, Dordrecht, Dordrecht, Netherlands, pp. 57–81, 2015.
- [5] S. Yu and Y. Kwon, "Development of VTOL drone for stable transit flight," *Journal of Computer and Communications*, vol. 5, no. 7, 2017.
- [6] Ö. Dündar, M. Bilici, and T. Ünler, "Design and performance analyses of a fixed wing battery VTOL UAV," *Engineering Science and Technology, An International Journal*, vol. 23, 2020.
- [7] H. Wu, Z. Wang, B. Ren et al., "Design and experiment OF a high payload fixed wing VTOL UAV system for emergency response," *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLIII-B3-2020, 2020.

- [8] A. M. D. Oca and G. Flores, "Unified controller for take-off and landing for a fixed-wing aircraft," in *Proceedings of the 2020 International Conference on Unmanned Aircraft Systems (ICUAS)*, Athens, Greece, September 2020.
- [9] Y. Govdeli, A. T. Tran, and E. Kayacan, "Multiple modeling and fuzzy switching control of fixed-wing VTOL tilt-rotor UAV," in *Proceedings of the International Fuzzy Systems Association World Congress 2019*, Lafayette, LA, USA, June 2019.
- [10] S. Wu, *Flight Control System*, Beijing University of Aeronautics and Astronautics Press, Beijing, China, 2013.
- [11] Q. Quan, *Design and Control of Multi-Rotor Aircraft*, Electronic Industry Press, Beijing, China, 2018.
- [12] J. Han, *Active Disturbance Rejection Control Technique*, National Defense Industry Press, Beijing, China, 2008.
- [13] J. Han, "From PID to active disturbance rejection control," *IEEE Transactions on Industrial Electronics*, vol. 56, no. 3, 2009.
- [14] S. Shen and J. Xu, "Attitude active disturbance rejection control of the quadrotor and its parameter tuning," *International Journal of Aerospace Engineering*, vol. 2020, Article ID 8876177, 15 pages, 2020.
- [15] Z. Wang, H. Zhao, D. Duan, Y. Jiao, and L. I. Jianbo, "Application of improved active disturbance rejection control algorithm in tilt quad rotor," *Chinese Journal of Aeronautics*, vol. 33, no. 6, pp. 1625–1641, 2020.
- [16] B. L. Stevens, F. L. Lewis, and E. N. Johnson, *Aircraft Control and Simulation*, John Wiley and Sons, Hoboken, NJ, USA, 3rd edition, 2016.
- [17] S. Song, W. Wang, K. Lu, and L. Sun, "Nonlinear attitude control using extended state observer for tilt-rotor aircraft," in *Proceedings of the 27th Chinese Control and Decision Conference (2015 CCDC)*, Qingdao, China, May 2015.
- [18] J. Song, Y. Jin, and Z. Du, "Application of extended state observer in longitudinal flight control of UAV," *Flight mechanics*, vol. 26, no. 6, pp. 76–83, 2008.
- [19] L. Zhang, *Research on Aircraft Flight Attitude Control Based on ADRC*, Nanjing University of Aeronautics and Astronautics, Nanjing, China, 2015.
- [20] J. Li, *Research on Adaptive PID Controller of Vertical Take-Off and Landing UAV*, South University of Science and Technology, Shenzhen, China, 2019.
- [21] Z. Gao, "Scaling and bandwidth parameterization based controller tuning," in *Proceedings of the American Control Conference*, Denver, CO, USA, June 2003.
- [22] A. Lambregts, "Vertical flight path and speed control autopilot design using total energy principles," in *Proceedings of the Guidance and Control Conference AIAA 83-2239*, pp. 559–569, Gatlinburg, TN, USA, August 1983.
- [23] Q. Zhang and A. N. Jinwen, "Xiaogang Liu Research on flight path/speed decoupling control Method based on TECS of Aircraft," *Journal of Northwestern Polytechnical University*, vol. 6, no. 3, pp. 384–387, 2004.
- [24] Q. Zhang and A. N. Jinwen, "A new design method of decoupled control system based on total energy to control flight speed/track," *Journal of Aeronautics*, vol. 7, no. 4, pp. 389–392, 2004.
- [25] L. F. Faleiro and A. A. Lambregts, "Analysis and tuning of a 'Total Energy Control System' control law using eigenstructure assignment," *Aerospace Science and Technology*, vol. 3, no. 3, pp. 127–140, 1999.
- [26] J. Li, Y. Xie, and W. Wang, "Application of total energy control principle in terrain following Flight Control," *Flight mechanics*, vol. 25, no. 1, pp. 67–70, 2007.
- [27] J. Li, Z. Hou, and Y. Xu, "Landing flight/thrust control system based on total energy theory," *Flight mechanics*, vol. 28, no. 2, pp. 35–38, 2010.
- [28] S. Wu, W. Cai, Y. Shen, and S. Guo, "Research on aircraft total energy control system I - principle analysis and system design," *Journal of Aeronautics*, vol. 7, no. 14, pp. 355–361, 1993.
- [29] A. Li, X. Xu, K. Ji, and G. Yuan, "Research on longitudinal landing control of large aircraft based on energy," *Journal of Northwestern Polytechnical University*, vol. 2, no. 1, pp. 22–26, 2011.
- [30] K. Ji, W. Wang, A. Li, and C. Wang, "Research on lateral control method of aircraft based on energy," *Flight Mechanics*, vol. 2, no. 1, pp. 32–35, 2011.
- [31] S. Park, J. Deyst, and P. H. Jonathan, "A new nonlinear guidance logic for trajectory tracking," in *Proceedings of the AIAA Guidance, Navigation and Control Conference and Exhibit*, Providence, RI, USA, August 2004.
- [32] Y. Wang, C.T. Li, and W. Chen, "Control law design method of angular rate for podded expendable loitering unit," *Flight Dynamics*, vol. 36, no. 1, pp. 38–42, 2018.

Research Article

Adversarial Sample Detection with Gaussian Mixture Conditional Generative Adversarial Networks

Pengfei Zhang  and Xiaoming Ju 

School of Software Engineering, East China Normal University, Shanghai, China

Correspondence should be addressed to Xiaoming Ju; xmju@sei.ecnu.edu.cn

Received 15 May 2021; Revised 13 July 2021; Accepted 8 August 2021; Published 13 September 2021

Academic Editor: Jie Chen

Copyright © 2021 Pengfei Zhang and Xiaoming Ju. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

It is important to detect adversarial samples in the physical world that are far away from the training data distribution. Some adversarial samples can make a machine learning model generate a highly overconfident distribution in the testing stage. Thus, we proposed a mechanism for detecting adversarial samples based on semisupervised generative adversarial networks (GANs) with an encoder-decoder structure; this mechanism can be applied to any pretrained neural network without changing the network's structure. The semisupervised GANs also give us insight into the behavior of adversarial samples and their flow through the layers of a deep neural network. In the supervised scenario, the latent feature (or the discriminator's output score information) of the semi-supervised GAN and the target network's logit information are used as the input of logistic regression classifier to detect the adversarial samples. In the unsupervised scenario, first, we proposed a one-class classifier based on the semisupervised Gaussian mixture conditional generative adversarial network (GM-CGAN) to fit the joint feature information of the normal data, and then, we used a discriminator network to detect normal data and adversarial samples. In both supervised scenarios and unsupervised scenarios, experimental results show that our method outperforms latest methods.

1. Introduction

Deep neural networks (DNNs) have achieved high accuracy in many classification tasks, such as speech recognition [1], objection detection [2], and image classification [3]. Although these DNNs are robust to random noise, they can mislead the model and cause it to output erroneous predictions when inputting small perturbations that are hard for humans to detect. In many machine learning applications (for example, in novelty detection [4], autonomous vehicles [5], and banking systems [6]), this prediction uncertainty will significantly reduce the model's safety.

Several methods have been proposed to protect against DNN attacks. One such method relies on the adversarial training method by adding adversarial samples in the training phase [7]. This method is robust to a variety of adversarial attacks but is ineffective against certain other attacks. To guarantee that there is no adversarial perturbation to fool the neural network within a given range, a

more computationally demanding and provable defense is used, employing either integer programming approaches [8, 9] or satisfiability modulo theories [10]. The above-mentioned methods require a lot of calculations and special training procedures. However, when the parameters and structure of the neural network are fixed, neither of these methods can be used without modifying the neural network structure or retraining the neural network.

Adversarial sample detection is a good solution to the above problems. In the supervised scenario, most methods train a binary classifier to distinguish whether the sample is a normal sample or an adversarial sample. In 2018, Lee et al. [11] established a class-conditional Gaussian distribution in the intermediate layers of the pretrained network and distinguished adversarial samples using the Mahalanobis distance. Meanwhile, Ma et al. [12] proposed the local intrinsic dimensionality (LID) and experimentally proved that the LID can be employed to represent a test sample's characteristics. Both supervised learning methods use normal

samples to train the feature extractor in the training stage. In the testing stage, the test samples are input to the feature extractor to obtain feature data. Finally, the feature data are input to the supervised classifier to realize the detection of normal samples and adversarial samples.

In the unsupervised scenario, alternatively, we can consider the unsupervised detection algorithm of adversarial samples. In 2017, Xu et al. [13] proposed a method to detect adversarial samples by comparing the model's prediction on a given image input with its prediction on the compressed image input version. In 2019, Yang et al. [14] proposed a feature attribute map of the adversarial samples close to the classification boundary; this map was different from the feature attribute map of the true data. In 2019, Roth et al. [15] introduced a statistical test based on the change in feature representations and log odds under noise; this approach is called odds-testing. PouyaSamangouei et al. [16] used GANs to model real images' distribution and find a close model's output to a test image, which did not contain adversarial perturbations; this confirmed that the adversarial sample has a data distribution far away from that of the normal sample, and it inspired other researchers to train a normal sample classifier to fit the distribution of real data. In 2018, an interesting analysis [16] showed how adversarial samples are propagated through neural network layer features; in 2019, Joshua et al. [17] studied the first-order classifier by training a discriminator with a generative adversarial network. These works [16, 17] inspired the present work. Regarding supervised learning and partial supervision methods, our method is inspired by the approach of Lee et al. [11]. They established a class-conditional Gaussian model through the Mahalanobis distance: they calculated the Mahalanobis distance between the true data and the adversarial sample and found that the feature distribution of the true data and that of the adversarial sample were different. Grosse et al., [18] also showed that adversarial samples have different distributions from normal data. Considering this finding, we also study the feature distribution information of normal samples through a semisupervised GAN in the present article. There are differences in the feature distributions between real samples and adversarial samples when the adversarial samples are input to the generator.

Our approach is described as follows. We used this difference to detect adversarial samples and true data. For unsupervised learning, we are inspired by the method of Engelsma and Jain [17]; we also employed the GM-CGAN to study the hidden layers' features and label information of the true data in the hidden layers of the pretrained network for joint feature distribution. When the sample belongs to true data, the discriminator will output a high predicted value. Conversely, when the sample is an adversarial sample, the joint feature of the hidden layers' feature and the label does not conform to the true data's feature distribution, and the discriminator will output a relatively low predicted value. We highlight that we do not directly study the joint features of true data and labels but instead use the intermediate layers' features of the pretrained network.

Figure 1 shows the framework of the proposed detection method. We train individual semisupervised generative

adversarial networks to study normal data distributions in the pretrained target network's hidden layers. G_E is the semisupervised GAN's encoder structure, and G_D is the semisupervised GAN's decoder structure. In a supervised scenario, the latent feature of the semisupervised GAN and the target network's logit information are used as the input of the external classifier to detect the adversarial samples. latent_1 represents the latent vector of the first semisupervised GAN, and latent_i represents the last latent vector. In addition to using the latent information to do experiments, we also use the discriminator's output score information of the semi-supervised GAN for supervised training in our experiment. In the case of an unsupervised scenario, input features (i.e., the reconstruction error vector error_i in the final layer of the last block group in the pretrained target network, the latent vector latent_i in the final layer of the last block group in the pretrained target network, and the logit vector of the target network) are used as Gaussian mixture conditional generative adversarial networks to study the distribution of features. Besides, the label information of the target network is used as the GM-CGAN's conditional information.

Our method has universal applicability, and the samples are tested without modifying any of the model's structure. Experiments conducted on the DenseNet and ResNet network architectures show that among the recently proposed detection methods, our method obtains the highest detection rate in both supervised and unsupervised learning scenarios. Our method is better than the method utilizing the Mahalanobis distance [11] in a supervised scenario. In a partially supervised scenario, our method and the Mahalanobis distance method have similar performance. In an unsupervised scenario, our method performs similar to or better than the odds-testing [15] method.

2. Materials and Methods

2.1. Preliminaries. In this work, we describe the deep neural network first and then introduce the observation of adversarial samples in the neural network. Finally, we present the foundations of the GAN and CGAN.

2.1.1. Neural Network. The neural network solves the classification problem of class k , and the final output result is obtained by logit through the softmax function. The neural network consists of M layers h_m :

$$z_m = h_m(z_{m-1}), \quad \text{for } m = 1, \dots, M. \quad (1)$$

In the above formula, M represents the total number of layers of the neural network, z represents the output of the neural network, and h is the hidden layer of the neural network.

2.1.2. Observation of Adversarial Samples in Neural Networks. In this article, we further verify that the adversarial sample and the normal sample are feature distributions in the target network's hidden layers. The distribution of the adversarial samples' feature data and the distribution

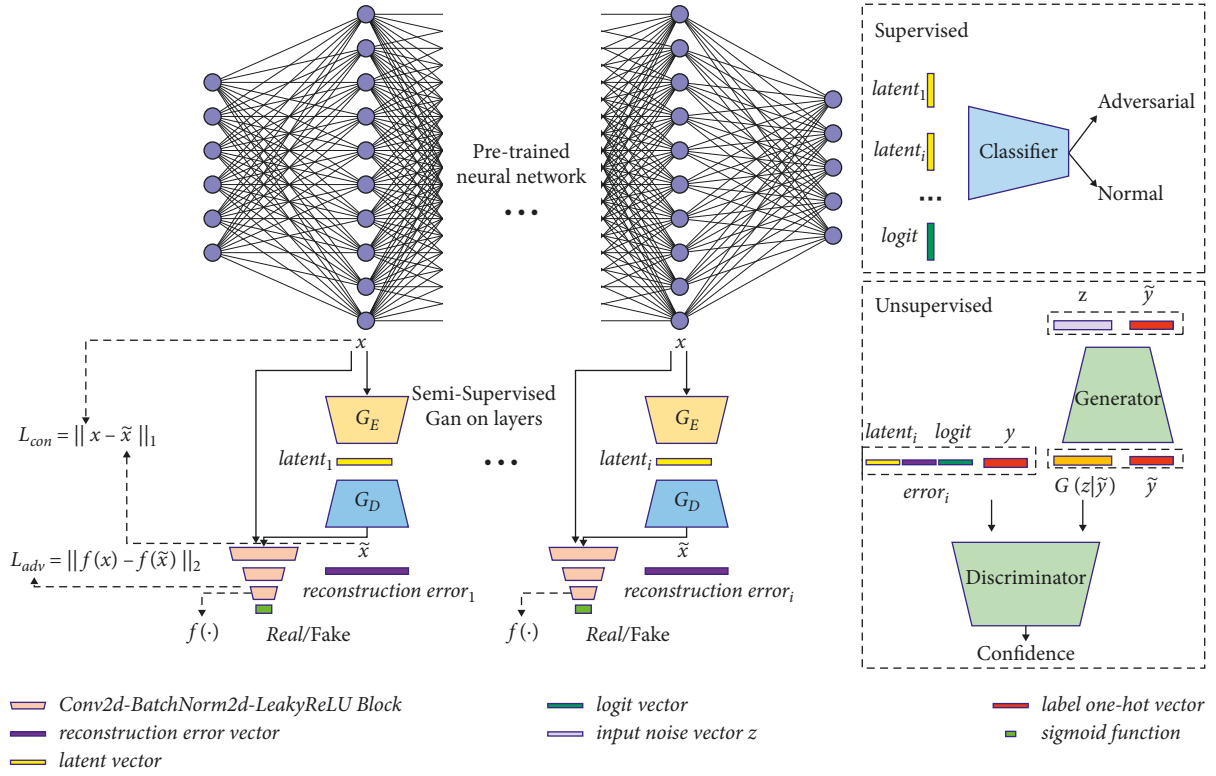


FIGURE 1: Framework of the proposed detection method.

of normal samples' feature data are different, and the influence of interference increases as the network deepens. Besides, adversarial samples will deviate from the real data.

The structure of the neural network contains nonlinear mappings and is formalized as follows:

$$F = h_1 \circ \dots \circ h_M, \quad (2)$$

where F is the final output of the neural network and h is the hidden layer of the neural network.

Since in practice, the Lipschitz constant $\text{lip}(h_i)$ of the neural network in each mapping is greater than 1 [19], we can assume that there is a small perturbation in the input space, and after the neural network propagates, the final layer has a vast representation distance. Our formula for the propagation of neural networks in high-dimensional space is formalized as follows:

$$\text{lip}(F) \approx \text{lip}(h_1) \dots \text{lip}(h_M), \quad (3)$$

where F is the final output of the neural network, h is the hidden layer of the neural network, and lip is the Lipschitz constant.

Due to the propagation properties of neural networks, as the number of neural network layers increases, the distribution of the adversarial samples will deviate farther from the distribution of normal samples. The difference between the real samples and the adversarial samples becomes more obvious in each sequent hidden layer.

2.1.3. GAN. The GAN [19] is relatively good for training generative models. It is composed of two adversarial modules: a generative model G used to describe the data distribution, and a probability discrimination model that determines whether the sample comes from the training data distribution instead of G . Both generator and discriminator are nonlinear mapping functions and have multilayer perceptrons.

To study the generator distribution p_g on the training data x , generator G builds a mapping function from the prior noise distribution $p_z(z)$ to the space of the generated data ($G(z; \theta_g)$). $D(x; \theta_D)$ (the output data of discriminator D) is a probability scalar used to determine the probability y that the data comes from the distribution of the training data. Generator G and discriminator D are trained at the same time, and we adjust their parameters so that the generator has minimal loss:

$$\log(1 - D(G(z))). \quad (4)$$

The loss of discriminator D is

$$\log(D(x)). \quad (5)$$

Thus, the generator and discriminator essentially maximize and minimize $V(D, G)$.

$$\begin{aligned} \min_G \max_D V(D, G) = & E_{x \sim p_{\text{data}}(x)} [\log(D(x))] \\ & + E_{z \sim p_z(z)} [\log(1 - D(G(z)))]. \end{aligned} \quad (6)$$

In equations (4) and (6), z is the input noisy data following the p_z distribution. Meanwhile, in equations (5) and (6), x is the input data following the p_{data} distribution. In equations (4)-(6), \log is a logarithmic function. The G is the generator network and the D is the generator network. E is the expectation of the distribution function. $p_{\text{data}(x)}$ is the distribution of real samples. $p_z(z)$ is a low-dimensional noise distribution. V is the loss function.

2.1.4. CGAN. If both generator and discriminator are conditioned with some additional information, the GAN can be extended to its conditional form. Y can be any type of auxiliary information, such as a class label or data of other forms. We can adjust $[x]$ using a discriminator and generator with y input as an additional input layer. In this way, the maximum and minimum objective function is as follows:

$$\begin{aligned} \min_G \max_D V(D, G) = & E_{x \sim p_{\text{data}}(x)} [\log(D(x|y))] \\ & + E_{z \sim p_z(z)} [\log(1 - D(G(z|y)))]. \end{aligned} \quad (7)$$

Here, z is the input noisy data following the p_z distribution, x is the input data following the p_{data} distribution, and y is the one-hot encoding of labels. \log is a logarithmic function. G is the generator network, and D is the generator network. E is the expectation of the distribution function. $p_{\text{data}(x)}$ is the distribution of real samples. $p_z(z)$ is a low-dimensional noise distribution. V is the loss function.

2.2. Related Work. We introduce adversarial attacks and adversarial defense in this work.

2.2.1. Adversarial Attacks. Adversarial attacks can be roughly divided into poisoning at training time or test time, and evasion. Adversarial attacks at training time are mainly conducted by adding maliciously tampered data into the training dataset during training so that the DNNs enter a suboptimal state, resulting in a decrease in model performance; this is called poisoning.

Meanwhile, evasion attacks involve tampering with the trained model's input, making the final prediction of the model incorrect. In both types of adversarial attack, an adversarial input modifies the other inputs in such a way that humans do not perceive the changes, but the DNNs make an incorrect final prediction.

For example, we add some minimal perturbations to pixels of the digit 2 in MNIST data so that the predicted value of the digit becomes 7, even though the digit still looks like the digit 2.

In this article, we study an evasion attack at test time. Given a test input x from class c , the adversarial attack aims to create the smallest perturbations so that the model's output will eventually become a specific class c' (targeted attack) or a class outside of class c (untargeted attack). This is formalized as an optimization problem, and its general form is as follows:

$$\min \|\delta\|_p \text{ s.t.}, \quad (8)$$

$$\hat{C}(x + \delta) = c' \text{ (targeted)}, \quad (9)$$

$$\text{or } \hat{C}(x + \delta) \neq c \text{ (untargeted)}. \quad (10)$$

In equations (9) and (10), \hat{C} is the trained model classifier. δ is the adversarial perturbation. x is the input data of the trained model. Based on the above general formula, many adversarial attack methods have been proposed; well-known methods include the fast gradient symbol algorithm (FGSM) [20], projected gradient descent (PGD) attack [22], Carlini-Wagner (CW) attack [23], DeepFool attack [24], and BIM attack [25]. These attack methods can be categorized into black-box attack or white-box attack methods depending on the extent of their knowledge about the DNNs classifier's parameters, structure, loss function, and algorithm. The most commonly used deep neural network attack methods are white-box methods because they assume a complete understanding of the system.

2.2.2. Defenses against Adversarial Attacks. Defending against neural networks is much more complicated than attacking them. Here, we summarize some current defense methods.

(1) *Adversarial Training.* It is a method that trains a better classifier to defend against adversarial samples. [25]. This is a method to add adversarial sample information in the training process of neural network classifier. For instance, one can add adversarial examples to the training data [26] for data augmentation [27] or add adversarial targets to the classification targets [28] for regularization [29]. Although the method is promising, it is difficult to determine which kind of attack is more suitable for the training way and how important the training way is; these problems are still unresolved [30].

(2) *Defensive Distillation.* The defensive distillation [30] training classifier makes it almost impossible for gradient-based attacks to directly generate adversarial samples on the trained network. This method uses the distillation training technique and hides the gradient between the logits and the output of the softmax function [26]. However, attacks can bypass this defense through the following three ways: (1) choosing an appropriate loss function, (2) directly calculating the gradient of the previous layer of the softmax layer instead of the gradient of the postlayer of the softmax layer, and (3) first attacking other vulnerable network models and then migrate to the trained distillation network.

(3) *Detecting Adversarial Samples.* Detecting adversarial samples can use statistical feature [29] methods or a classification network [30] to achieve defense. We build different detection classifiers for different attack methods to determine whether the input is a normal sample. The detector uses normal samples and adversarial samples for training. When the

training and testing adversarial samples are generated from the same way, and the adversarial perturbation is obvious enough, the detector shows good performance. However, this means of defense cannot be well generalized to different attack.

Our method is inspired by the approach of Lee et al. [11]. They established a class-conditional Gaussian model through the Mahalanobis distance: they calculated the Mahalanobis distance between true data and the adversarial sample and found that the feature distribution of the true data and that of the adversarial sample were different. Our method takes advantage of the difference between the adversarial samples and the normal samples in the middle layers of the neural network. In 2019, Joshua et al. [17] studied the first-order classifier by training a discriminator with a generative adversarial network. In our unsupervised method, we use the feature vectors extracted from the middle layer to train the discriminator network. Because our GM-CGAN is used to train a one-class classifier on the microfeature distribution of the middle layers of the pretrained network, we capture the weak perturbation of adversarial samples. Thus, our method will have better detection performance than the three types of methods described above.

2.3. Our Approach. In this work, we introduce the semi-supervised GAN's structure and training method for studying the hidden layers' feature distribution of the pretrained network first. Then, we present our detection methods under supervised, partially supervised, and unsupervised scenarios.

2.3.1. Semisupervised GAN. Samet Akcay et al. [28] inspired us to use the semisupervised anomaly detection structure. They used an encoder-decoder-encoder structure to study the data distribution of the input image. For simplicity, we used an encoder-decoder structure in the generator part of the semisupervised GAN to analyze the feature distribution of the target network's hidden layers.

The formal principle behind the semisupervised GAN is as follows. Generator G first reads the intermediate layer feature x of the target network, where $x \in R^m$, and forward passes it to the encoder network G_E . With the use of convolutional layers followed by batch norm and ReLU() activation, G_E downscales x by compressing it to a latent vector, where latent $\in R^d$ (d represents the best dimension). In our experiment, d is set to 128. The decoder network G_D of generator G is composed of convolution transpose layers, the batch-norm function, and the ReLU() activation function. Finally, the latent vector is input to G_D to reconstruct the intermediate layer feature \tilde{x} of the target network. Finally, generator G generates the intermediate layer feature \tilde{x} via $\tilde{x} = G_D(\text{latent})$, where latent = $G_E(x)$.

(1) Adversarial Loss. Because the GAN is unstable during the training phase, we add feature matching loss to the training phase. In the ordinary GAN's training phase, generator G is updated based on discriminator D . Following the work of Salimans et al. [29], feature matching is employed to reduce instability during training. We update this approach based on the internal representation of discriminator D . First, we assume that there is an f function of the intermediate layer of

discriminator D . For the input data x that satisfies the p_x distribution and outputs the intermediate layer features of discriminator D , the feature matching loss calculates the L_2 distance between the original feature and the generated feature. The form of our formalized adversarial loss L_{adv} is as follows:

$$L_{adv} = E_{x \sim p_x} \|f(x) - E_{x \sim p_x} f(G(x))\|_2. \quad (11)$$

In equation (11), x is the input data; f is the function of the intermediate layer of discriminator D , G is the generator network, D is the discriminator network, and p_x is the distribution of real samples. The adversarial loss L_{adv} is the L_2 loss. E is the expectation of the distribution function.

(2) Contextual Loss. While adversarial loss makes the generated adversarial samples deceive discriminator D , there is only one adversarial loss, and the generator cannot be optimized according to the input data's context information. Punishing the generator by measuring the distance between the input data and the generated data can remedy this problem. Isola et al. [30] showed that the fuzzy results due to L_1 loss are less than those due to L_2 loss. Therefore, we penalize G by measuring the L_1 distance between the original input data and the generated data $\tilde{x} = G(x)$. Thus, the contextual loss L_{con} is formalized as follows:

$$L_{con} = E_{x \sim p_x} \|x - G(x)\|_1. \quad (12)$$

In equation (12), x is the input data, E is the expectation of the distribution function, and p_x is the distribution of real samples. The contextual loss L_{con} is the L_1 loss.

In this way, the generator encodes normal data but cannot encode adversarial samples because our generator G and discriminator D are optimized for normal data. The loss function we trained is as follows:

$$L = W_{adv} L_{adv} + W_{con} L_{con}, \quad (13)$$

where W_{adv} is the weight coefficient of L_{adv} and W_{con} is the weight coefficient of L_{con} . W_{adv} and W_{con} are both positive integers; $W_{adv} = 1$, and $W_{con} = 15$. L_{adv} is the adversarial loss in equation (11). L_{con} is the contextual loss in equation (12). The semisupervised GAN training flow chart is described in Algorithm 1.

2.3.2. Supervised Scenario and Partially Supervised Scenario. We used the semisupervised GAN to study the normal data's feature distribution information in the hidden layers of the pretrained network. Then, we input the normal data and the adversarial samples into the pretrained network and obtain the corresponding latent features through the semisupervised GAN. Finally, the features are input into the supervised classifier logistic regression classifier to realize supervised classification and partially supervised classification. The supervised scenario and partially supervised scenario training flowchart is described in Algorithm 2.

2.3.3. Unsupervised Scenario. Due to the discrepancy between the feature distributions of the normal data and the adversarial samples in the hidden layers of the pretrained

Input: train sample x into the pretrained target network.
for each layer $l \in 1, \dots, L$ do
 Train individual semisupervised generative adversarial networks in layer l .
 A semisupervised GAN_l is obtained.
end for
return semisupervised GAN_l $l \in 1, \dots, L$

ALGORITHM 1: We train individual semisupervised generative adversarial networks to study normal data distributions in the pretrained target network's hidden layers.

Input: the normal data x and the adversarial samples x_{adv} into the pretrained network.
for each layer $l \in 1, \dots, L$ do
 Input: the hidden layer l into semisupervised GAN_l (Algorithm 1)
 latent $_l$ is obtained.
 error $_l$ is obtained.
end for

The normal data x latent feature is $\sum_l \text{latent}_x$ and the logit vector of the target network.

The adversarial samples x_{adv} latent feature is $\sum_l \text{latent}_{x_{\text{adv}}}$ and the logit vector of the target network.

We use the normal data x latent feature and adversarial samples x_{adv} latent feature input the support vector machine classifier.
 return the evaluation data's AUROC.

ALGORITHM 2: In the supervised and partially supervised scenario, the final detection classifier is the support vector machine classifier. We used 10% of the test set as training data and 90% as evaluation data.

network, we use the hidden layers' difference features of the normal data and its label as the joint feature information for training the CGAN and a good discriminator.

The loss function of the target is as follows:

$$\begin{aligned} \min_G \max_D V(D, G) = & E_{x \sim p_{\text{normal}}(x)} [\log(D(x|y))] \\ & + E_{z \sim p_z(z)} [\log(1 - D(G(z|y)))] \end{aligned} \quad (14)$$

Here, in order to reduce the amount of calculation, x is the joint feature data that includes the reconstruction error vector in the final layer of the last block group in the pretrained target network, the latent vector in the final layer of the last block group in the pretrained target network, and the logit vector of the target network. p_{normal} is the distribution of real samples, and p_z is the low-dimensional noise distribution; x follows a normal distribution p_{normal} , and z is the input noisy data following the p_z distribution, and y is the one-hot encoding of labels. Usually, $U[-1, 1]^d$ or multivariate normal distribution information $N(0, I_{d \times d})$ are used as the noise input during the GAN's training. G is the generator network and D is the discriminator network. We emphasize that to better study the joint feature distribution of normal data and labels, we use the inherent multimodal distribution feature of p_x . Its specific form is as follows:

$$p_z(z) = \sum_{k=1}^K \alpha_k * p_k(z). \quad (15)$$

Here, K is defined as the number of Gaussian distributions in the mixture model, which is the number of neural network block groups in our experiment. $p_k(z)$ is defined as

multivariate normal distribution information $N(\mu_k, \sum k)$, where $\forall k \in [K]$, $\alpha_k = 1/K$. z is the input noisy data. For data evaluation, we used the normal sample's one-hot encoding of labels and the adversarial sample's one-hot encoding of labels as the trained GM-CGAN's conditional information during the testing phase. The supervised scenario and partially supervised scenario training flowchart is described in Algorithm 3.

3. Results and Discussion

3.1. Experiments. We test our detection method against DeepFool, FGSM, BIM, PGD, and CW adversarial attacks on CIFAR10 [32], CIFAR100 [33], and SVHN [34] datasets. We used the ResNet-34 [35] and DenseNet-BC [36, 37] models. Similar to Lee's method, we chose to train the semisupervised GAN on the last layer of the basic block of the two neural network models; then, we extracted the hidden layers' features from the target network. For convenience of calculation, if the feature shape of the dataset in the network's hidden layers is the same, we used the same semisupervised GAN.

In the supervised and partially supervised scenario, the final detection classifier is the logistic regression classifier; we used 10% of the test set as training data and 90% as evaluation data. In the unsupervised scenario, we used the GM-CGAN as the final detection classifier trained on the training samples that not include adversarial samples and noise samples. Our analyses of reconstruction error and L2 norm are presented in Figures 2–4. When we train the GM-CGAN, we select the input features as the reconstruction error vector in the final layer of the last block group in the pretrained target network, the latent vector in the final layer of the last block group in the pretrained target network, and

Input: the normal data x into the pretrained network.
for each layer $l \in 1, \dots, L$ do
 Input: the hidden layer l into semisupervised GAN $_l$ (Algorithm 1)
 latent $_l$ is obtained.
 error $_l$ is obtained.
end for

In order to reduce the amount of calculation, the GM-CGAN's input features are the reconstruction error vector in the final layer of the last block group in the pre-trained target network, the latent vector in the final layer of the last block group in the pre-trained target network, and the logit vector of the target network. The GM-CGAN's label information is the label of the output of this neural network model.

return the evaluation data's AUROC.

ALGORITHM 3: In the unsupervised scenario, we used 10% of the test set as training data and 90% as evaluation data.

the logit vector of the target network. The GM-CGAN's conditional information is the label of the output of this neural network model.

The features we input are important for detecting adversarial samples. In the study by Yang et al. [19], the reconstruction error vector of the L1 norm was shown to reflect the discrepancy between the given sample and the real sample. The latent feature vector's norm reflects whether a given sample can be generated on the data manifold. In Figure 2, we can also view the importance of the logit norm. This can reduce the computational complexity and allows us to better capture the difference information. Similar to Lee's [11] method, we use logistic regression classifier. The hyperparameters of the SVM classifier are fine-tuned.

First, similar to how Yang et al. [14] approached the problem, we analyzed the norm and reconstruction error information of the hidden layer's latent feature vectors generated from the semisupervised GAN (Figures 2 and 3). We also analyzed the joint feature information of the norm and reconstruction error (Figure 4). Finally, we assessed the importance of different hidden layers' features (Figure 5). Here, due to limited space of article, we only present figures for the ResNet model under attack by five methods, on the CIFAR10 dataset. This analysis puts forward a strong argument that the reconstruction error and the latent norm can fully explain the data manifold, which can help us to detect adversarial samples.

For the sake of fairness, our method and the method of Lee et al. [11] are initialized with the same settings. In the experimental test stage of Lee et al., noisy data and adversarial samples are generated for normal test data. Our semisupervised GAN obtains latent features, and then, we used these latent features to train the supervised and unsupervised classifiers. We used 10% of the test set as training data and 90% as evaluation data and performed five-fold cross-validation.

3.2. Result. In Figure 2, we visualized the distribution information of normal (green) and adversarial (red) samples through kernel density estimation in 2D space. In the subfigure, x -axis represents the L2 norm in the latent vector produced by the generator and y -axis represents probability density. These figures are generated for the CIFAR10 dataset, with the ResNet model. The adversarial samples flow through 5 semisupervised

GANs and the target network's logit layer, and hence, there are 6 rows of subfigures. The difference between the real sample and the adversarial sample is obvious in the last few layers of the neural network, especially in the logit layer.

In Figure 3, we visualized the distribution information of normal (green) and adversarial (red) samples through kernel density estimation in 2D space. In the subfigure, x -axis represents reconstruction error and y -axis represents probability density. These figures are generated for the CIFAR10 dataset, with the ResNet model. The adversarial samples flow through 5 semisupervised GANs, and hence, there are 6 rows of subfigures.

In Figure 4, we visualized the distribution information of normal (green) and adversarial (red) samples through kernel density estimation in 2D space. In the subfigure, x -axis represents the L2 norm in the latent vector produced by the generator and y -axis represents reconstruction error. These figures are generated for the CIFAR10 dataset, with the ResNet model. The adversarial samples flow through 5 semisupervised GANs, and hence, there are 6 rows of subfigures.

Figure 5 presents the AUROC of the threshold-based detector using the latent vector generated from the semisupervised GAN's generator at different basic blocks of ResNet trained on CIFAR10 dataset and the logit vector of the target network. We measured the detection performance using adversarial samples produced by FGSM, BIM, DeepFool, CWL2, and PGD. We also measured the detection performance using the entire latent vectors generated from the semisupervised GAN's generator (see the last subfigure, i.e., Feature Ensemble).

To generate the results for the supervised scenarios (Table 1), the final detection classifier is the support vector machine classifier whose input includes all latent vectors. For the partially supervised scenarios (Table 2), the final detection classifier is the support vector machine classifier whose input only includes FGSM samples.

Meanwhile, for the unsupervised scenarios (Table 3), we used the GM-CGAN as a one-class classifier whose input features are the reconstruction error vector, the latent vector in the final layer of the last block group (obtained by the semisupervised GAN), and the logit of the target network. The GM-CGAN's conditional information is the label of the target network. We train the classifier based on the training data, which does not include any adversarial samples and noisy samples.

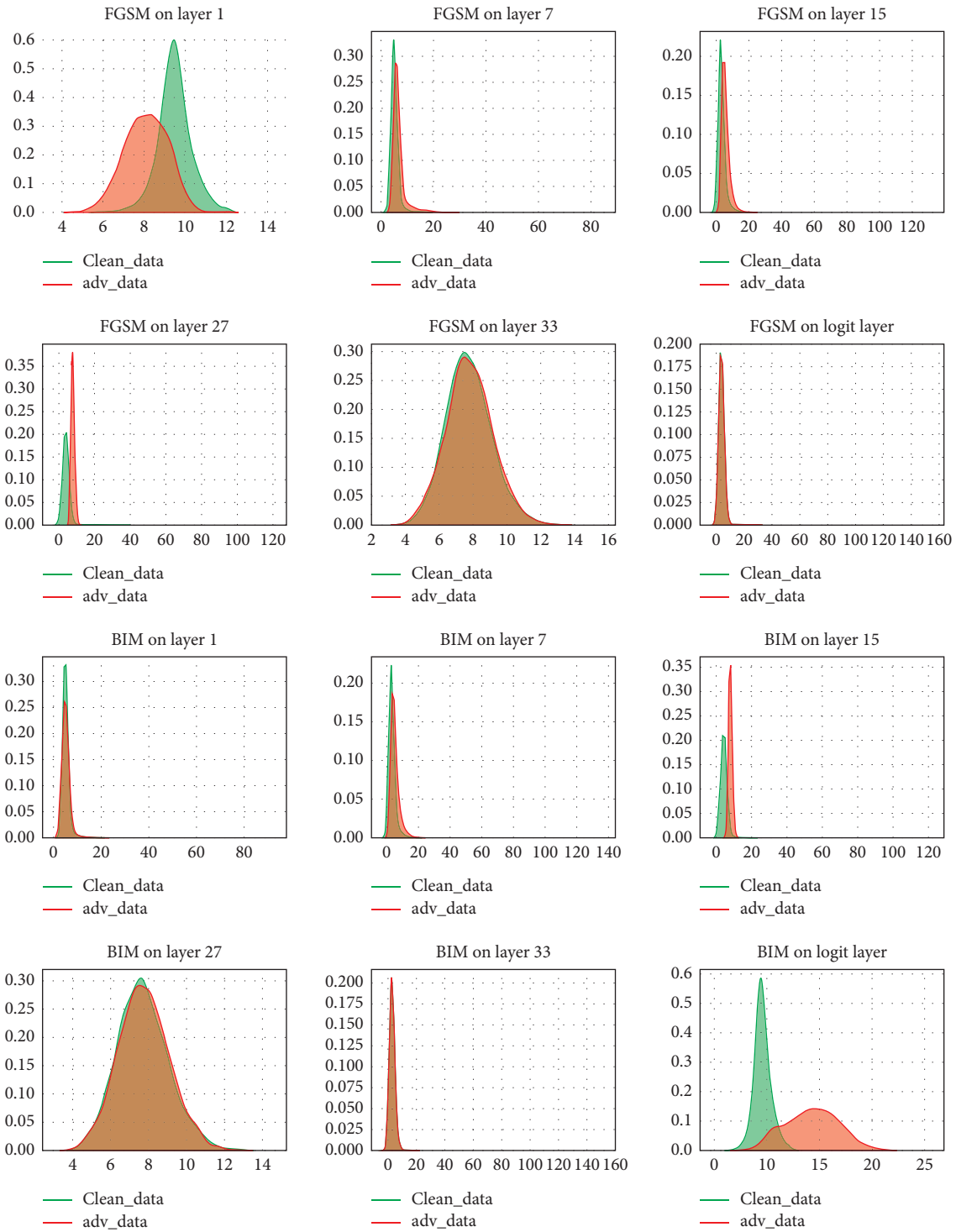


FIGURE 2: Continued.

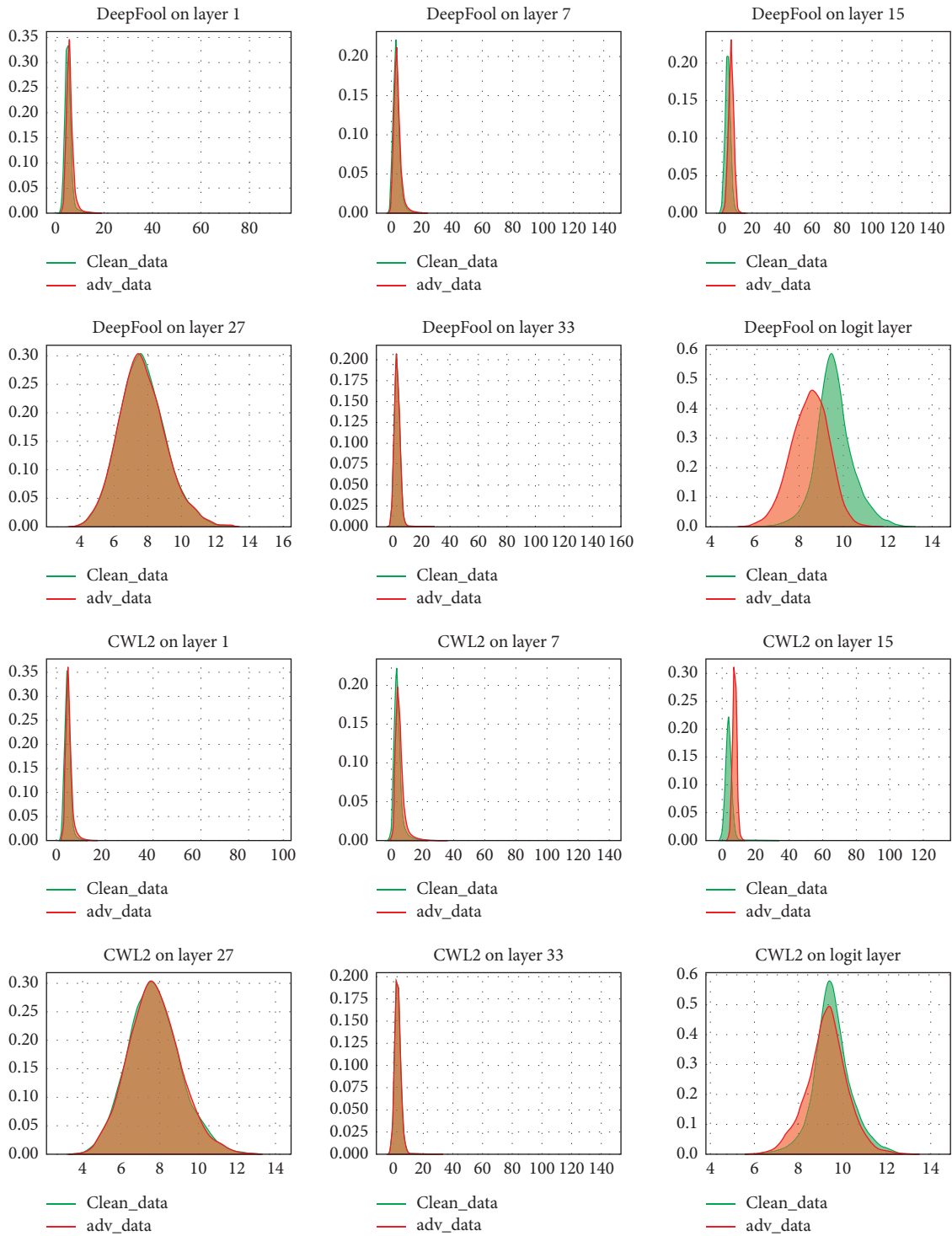


FIGURE 2: Continued.

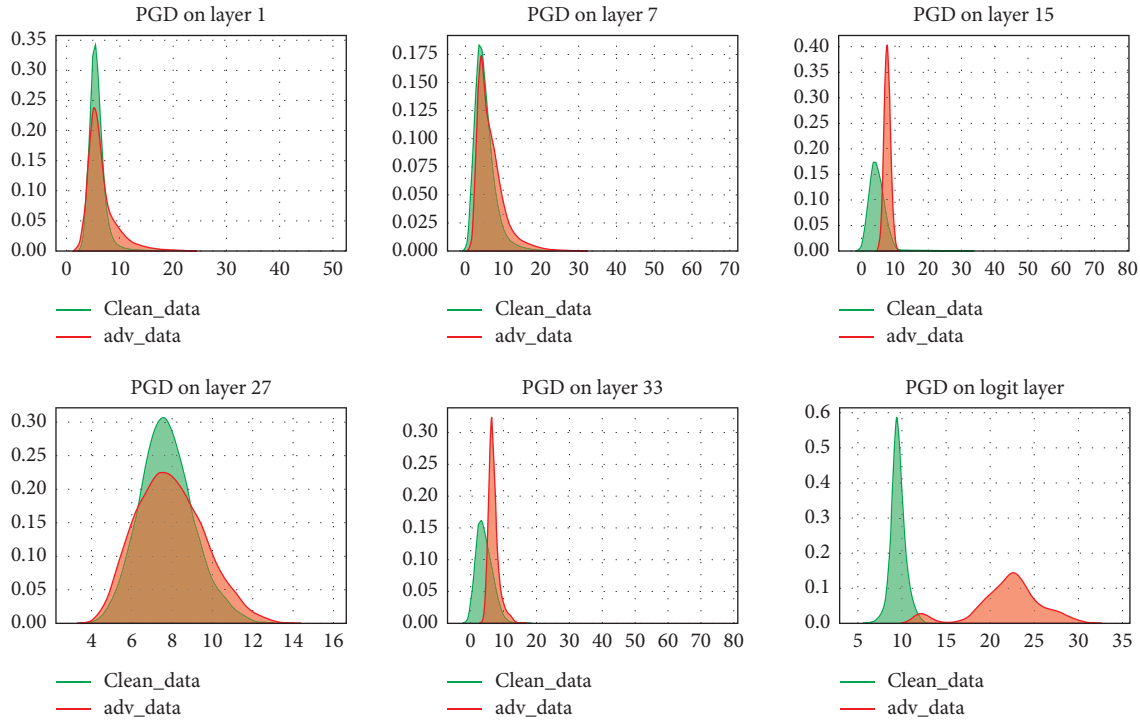


FIGURE 2: Analysis of feature norms of different network layers.

3.3. Discussion. In our experiment, we utilized the PGD-100 attack. In the supervised scenario, our method performed significantly better than that of Lee et al. Our method reduced the preprocessing time for input data and reduced the total amount of calculations. We note that the Mahalanobis distance covariance matrix must be full rank and that it cannot handle problems on nonlinear manifolds. In the deep layers of the neural network, the Mahalanobis distance cannot provide a reliable measure of the distance between the data, which mainly exist in a nonlinear form; ultimately, the Mahalanobis distance is unstable in such situations. The source of this instability is the covariance matrix. In Lee's method, the distance between data of different network blocks is very small compared with the distance between high-dimensional adversarial data and the real data. Furthermore, Lee's method cannot reflect the difference between the adversarial data and the real data. In contrast, our semisupervised GAN method maps different network block feature data to low-dimensional space, thereby obtaining more sample feature information. Moreover, it can overcome the drawback of the Mahalanobis distance being unsuitable for the determination of nonlinear data. Considering these key characteristics, it is clear why our method achieved the best results in the supervised classification task.

We highlight that our method improved the detection of DeepFool to above 94.68%.

For the partially supervised scenario, we used the logistic regression classifier as the final classifier with FGSM samples. Although it did not achieve the same effect as Lee's method in this scenario, our method still achieved good results. Here, we take the ResNet model and the CIFAR10 dataset as an example (Table 2). For BIM attacks, the detection AUROC dropped from 98.91% to 73.19%; meanwhile, for FGSM attack detection, the AUROC was 99.98%. We argue that there might be a trade-off between performance on a fully supervised scenario (where our method had an AUROC close to 100% in some cases) and an ability to generalize to other attacks. In Figure 4, we find that BIM attacks and FGSM attacks deviate from the real data and have great inconsistencies; this is mainly manifested in the 1st, 7th, and 27th layers of the network. By the same token, the ways in which different attacks deviate from the true data are also different. Therefore, only partial supervision is suitable for FGSM attacks.

For the unsupervised scenario, our method performed better than the odds-testing method [15], except for with the PGD-100 attack. We argue that not all attack differential features are present in the last layer. The PGD attack consists of initializing the search for an adversarial

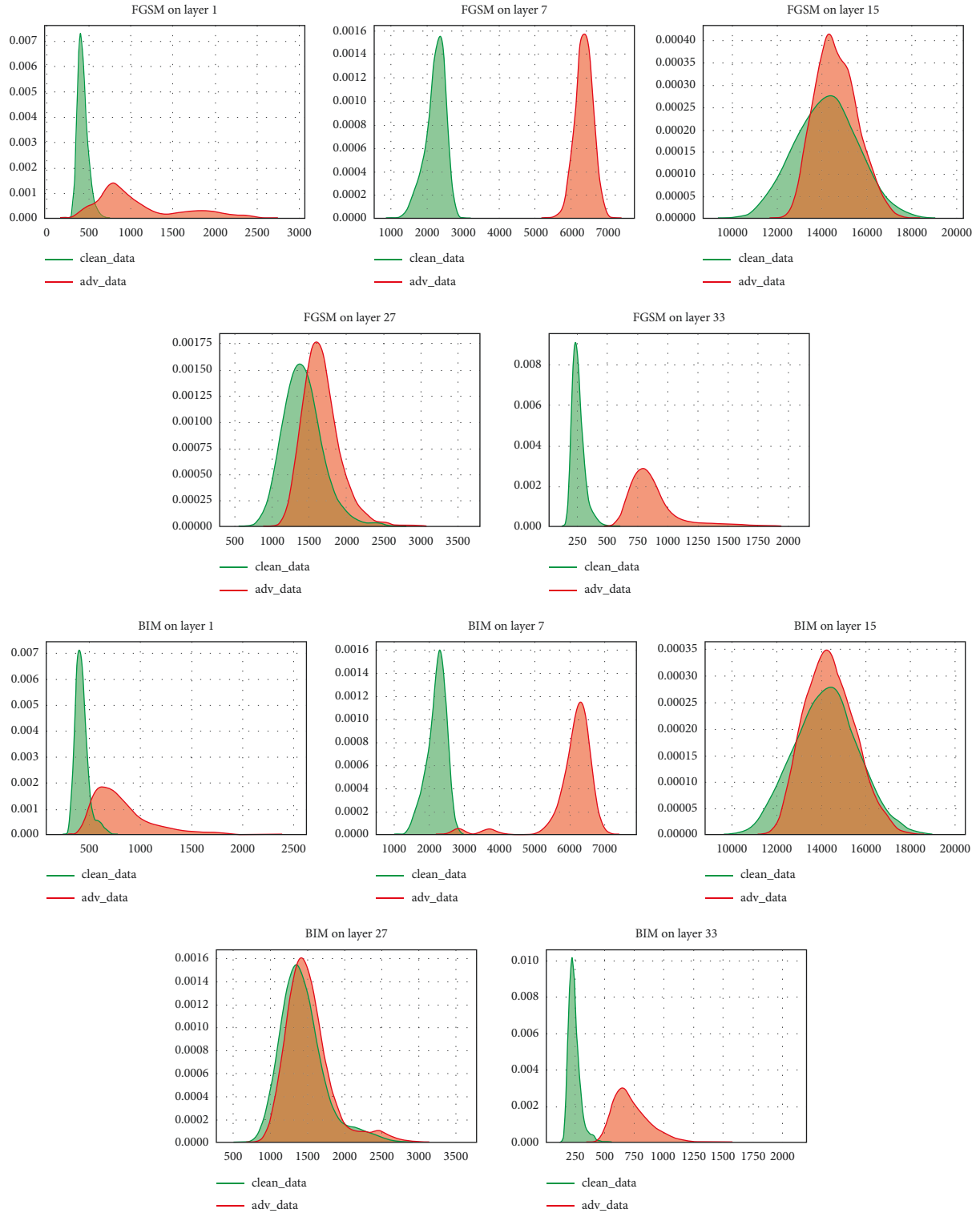


FIGURE 3: Continued.

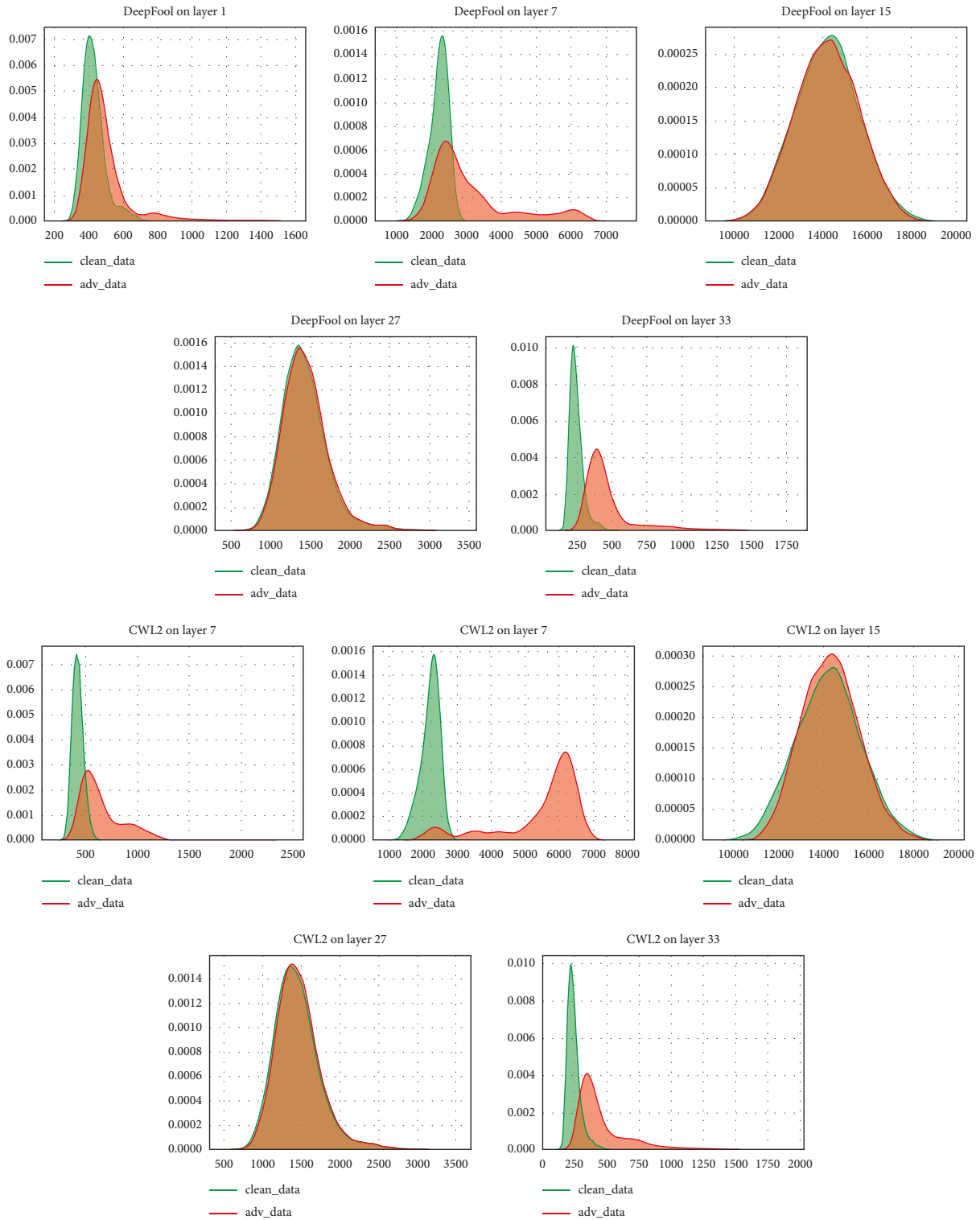


FIGURE 3: Continued.

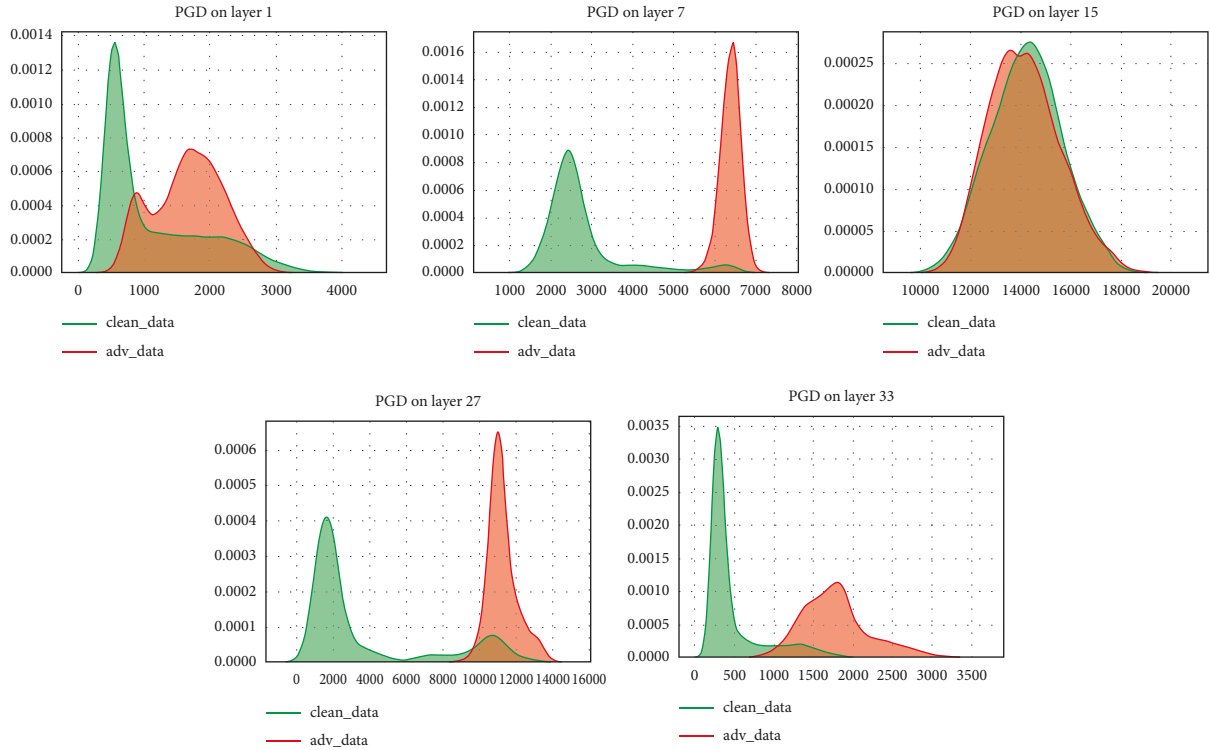


FIGURE 3: Analysis of reconstruction errors in different network layers.

example at a random point within the allowed norm ball, then running several iterations of the basic iterative method. The PGD-100 has stronger attack performance because of the basic iterative method. We think the performance feature of PGD’s adversarial samples in the network middle layer is more similar to those of real samples. The input at the end of the network is different from the real one. The PGD-100 can attack deep neural networks, such as ResNet-34 and DenseNet, which contain many features. GM-CGAN contains fewer features and is easier to be attacked by PGD-100. Because GM-CGAN is also a classifier based on neural network, the features in the hidden layers space will also be attacked by PGD. It reduces the defensive performance of GM-CGAN. For example, as shown in Figures 3 and 4, the most obvious difference in the PGD-100 attack is in the 15th layer. Although our method requires multiple forward propagations of neural networks like the odds-testing method, we provide a new idea for detecting adversarial samples: a new one-class classifier. We emphasize that our one-class classifier does not require any noisy data during the training process compared with the odds-testing method,

and the training method is simple and easy to operate. Additionally, the number of forward propagations of our method is relatively small. Through Figures 2–4, we find that as the network layer deepens, the data distribution of the adversarial sample deviates farther from the characteristic distribution of the normal data. The L2 norm information of the logit can better reflect this difference than the latent vector’s norm information. Compared with the L2 norm information, the reconstruction error can reveal the distribution discrepancy between the adversarial sample and the normal sample earlier, and the deeper the network layer, the greater the discrepancy. The combined information of the reconstruction error and L2 norm can also reflect this trend. In general, the discrepancy between the distribution of real samples and adversarial samples is more obvious in the last few layers. In Figure 5, we analyzed the detector’s performance with different basic blocks (which have different latent vector characteristics). We also analyzed the performance of the detector after the integration of different basic block features. Like in Figures 2–4, in most instances, the discrepancy is most obvious in the last layer. At the same time, we found that

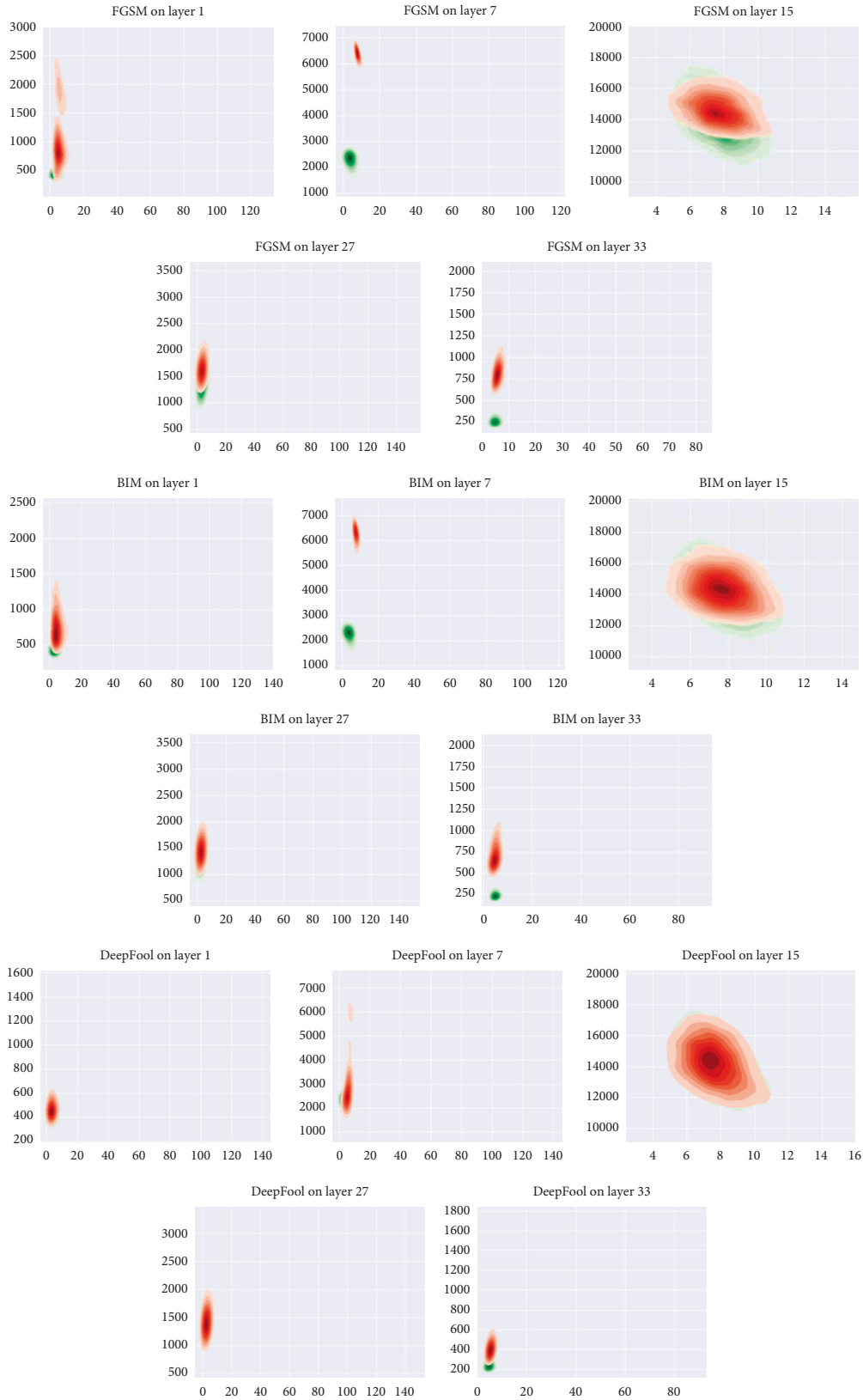


FIGURE 4: Continued.

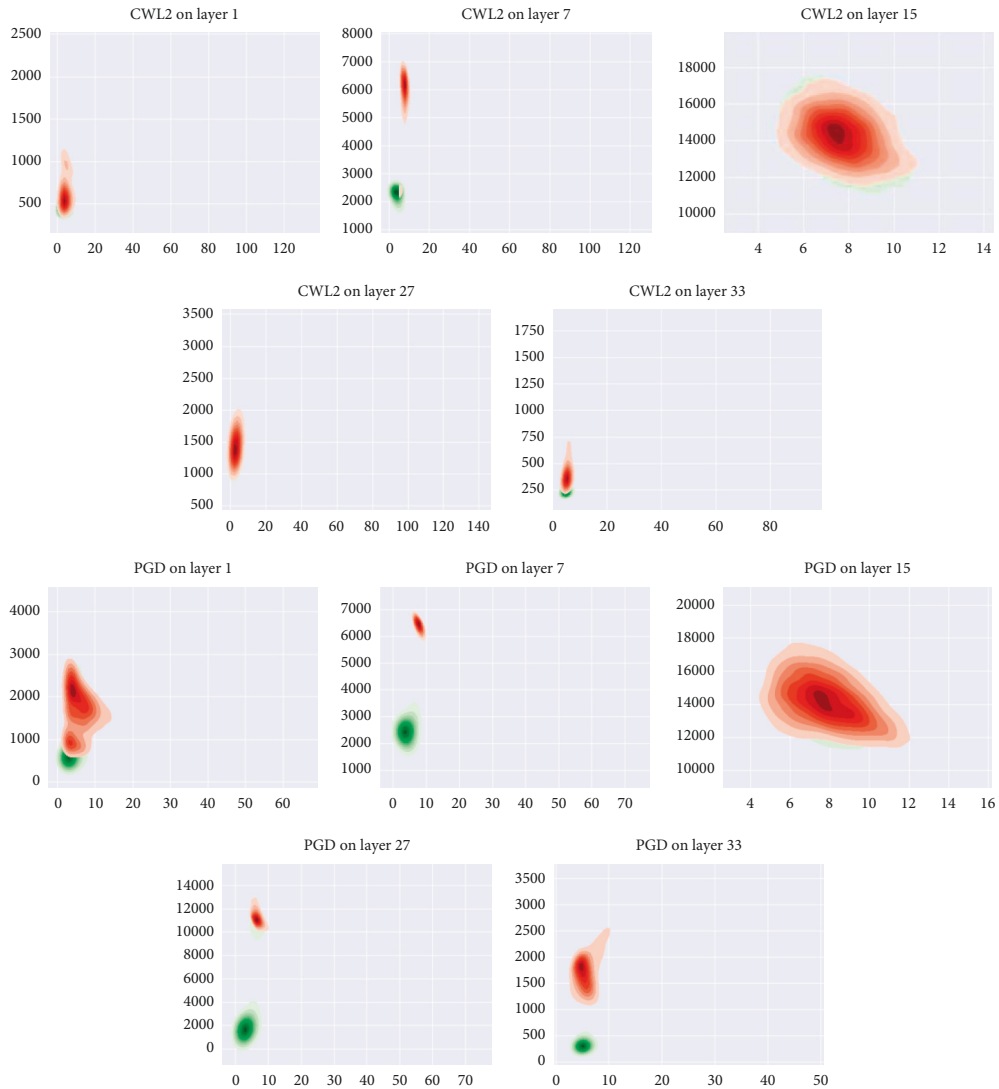


FIGURE 4: Analysis of reconstruction errors and latent norms in different network layers.

the logit vector is significant for the detection of adversarial samples. In our supervised and partially supervised experiments, we ensemble the latent feature vectors; we believe that this treatment can provide adaptability to different adversarial attack strategies and lead to good performance.

This method has practical significance. For example, this method can be used in target recognition. Without modifying the original neural network, this method can detect images with adversarial perturbation. For example, if adversarial perturbation is added to a picture of a kitten, it may be recognized as other animals in target recognition.

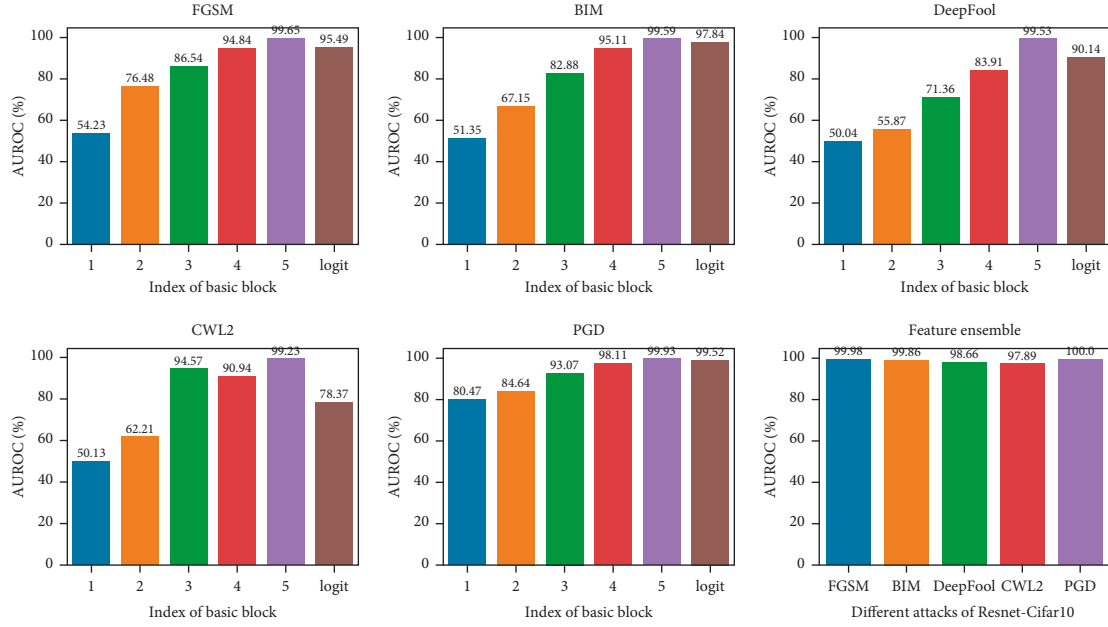


FIGURE 5: AUROC (%) of the threshold-based detector using the latent vector generated by semisupervised GAN's generator at different basic blocks of ResNet trained on the CIFAR10 dataset and the logit vector of the target network. (a) FGSM. (b) BIM. (c) DeepFool. (d) CWL2. (e) PGD. (f) Feature ensemble.

TABLE 1: Supervised scenarios for detecting adversarial samples.

Model	Dataset	Method	Supervised scenario				
			FGSM	BIM	DeepFool	CW	PGD
DenseNet	CIFAR10	Mahalanobis	99.94	99.78	83.41	87.31	97.79
		Ours (latent)	99.95	99.98	97.46	96.11	99.42
		Ours (score)	99.98	99.93	98.54	96.13	99.34
	CIFAR100	Mahalanobis	99.86	99.17	77.57	87.05	79.24
		Ours (latent)	100.00	99.86	97.22	98.01	90.35
		Ours (score)	99.89	99.90	97.34	98.02	91.36
	SVHN	Mahalanobis	99.85	99.28	95.10	97.03	98.41
		Ours (latent)	99.95	99.85	99.25	98.65	99.49
		Ours (score)	99.90	99.80	99.35	98.23	99.12
ResNet	CIFAR10	Mahalanobis	99.94	99.57	91.57	95.84	89.81
		Ours (latent)	99.98	99.86	98.66	97.89	100.00
		Ours (score)	99.87	99.92	98.65	98.11	98.65
	CIFAR100	Mahalanobis	99.77	96.90	85.26	91.77	91.08
		Ours (latent)	99.92	99.11	94.68	97.21	99.95
		Ours (score)	99.93	97.91	94.34	92.34	92.34
	SVHN	Mahalanobis	99.62	97.15	95.73	92.15	92.24
		Ours (latent)	99.96	99.46	99.54	99.30	99.98
		Ours (score)	99.65	98.34	96.54	97.56	97.45

TABLE 2: Partially supervised scenarios in detecting adversarial samples.

Model	Dataset	Method	Partially unsupervised scenario				
			FGSM	BIM	DeepFool	CW	PGD
DenseNet	CIFAR10	Mahalanobis	99.94	99.51	83.42	87.95	81.84
		Ours (latent)	99.95	90.79	98.06	95.75	76.00
		Ours (score)	99.98	92.23	96.09	97.18	78.00
	CIFAR100	Mahalanobis	99.86	98.27	75.63	86.20	39.32
		Ours (latent)	100.00	89.86	83.14	79.08	62.35
		Ours (score)	99.89	90.15	80.19	81.09	64.15
	SVHN	Mahalanobis	99.85	99.12	93.47	96.95	81.40
		Ours (latent)	99.95	99.00	98.71	98.16	94.15
		Ours (score)	99.90	99.01	98.29	97.18	92.16

TABLE 2: Continued.

Model	Dataset	Method	Partially unsupervised scenario				
			FGSM	BIM	DeepFool	CW	PGD
ResNet	CIFAR10	Mahalanobis	99.94	98.91	78.06	93.90	100.00
		Ours (latent)	99.98	73.19	96.79	95.71	100.00
		Ours (score)	99.87	96.15	94.13	94.10	100.00
	CIFAR100	Mahalanobis	99.77	96.38	81.95	90.96	99.85
		Ours (latent)	99.92	81.18	83.32	86.63	100.00
		Ours (score)	99.93	80.10	80.13	87.01	100.00
		Mahalanobis	99.62	95.39	72.20	86.73	99.92
		Ours (latent)	99.96	74.89	95.97	89.65	99.96
		Ours (score)	99.65	75.14	95.10	89.75	99.23

TABLE 3: Unsupervised scenarios for detecting adversarial samples.

Model	Dataset	Method	FGSM	Unsupervised scenario			
				BIM	DeepFool	CW	PGD
DenseNet	CIFAR10	Odds-testing	45.23	69.01	58.30	61.29	97.93
		GM-CGAN	87.89	73.69	80.81	78.12	46.45
	CIFAR100	Odds-testing	43.22	65.22	49.53	47.64	96.91
		GM-CGAN	98.09	68.34	74.42	65.72	41.09
	SVHN	Odds-testing	56.14	71.11	67.81	70.71	99.25
		GM-CGAN	83.35	73.56	81.86	80.16	46.40
ResNet	CIFAR10	Odds-testing	46.32	59.85	75.58	57.58	96.18
		GM-CGAN	96.68	64.74	79.48	73.49	98.80
	CIFAR100	Odds-testing	38.26	43.52	61.13	44.74	93.73
		GM-CGAN	80.96	81.20	80.35	67.56	91.30
	SVHN	Odds-testing	65.09	70.31	77.05	72.12	99.08
		GM-CGAN	94.97	89.04	94.71	83.41	97.52

This method can be used to detect kitten images with anti-interference. This method can prevent the recognition image from making mistakes.

4. Conclusions

Our article intends to discover the adversarial samples in training data in order to prevent the generation of the highly overconfident distribution in the test phase. The proposed method designs a semisupervised generative adversarial network that is applied to the output of the hidden layers in a neural network to detect the variation of the adversarial samples without modifying the structure of the neural network. In the supervised scenario, the latent feature (or the discriminator's output score information) of the semisupervised GAN and the target network's logit information are used as the input of the external classifier logistic regression classifier to detect the adversarial samples. In the unsupervised scenario, first we proposed a one-class classifier based on the semisupervised Gaussian mixture conditional generative adversarial network (GM-CGAN) to fit the joint feature information of the normal data and then we used a discriminator network to detect normal data and adversarial samples. The novel contribution is that the output of hidden layers of a neural network is analyzed without modifying the neural network.

Data Availability

The [DATA TYPE] data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by Science and Technology Project of State Grid Zhejiang Electric Power Co., Ltd.

References

- [1] D. Amodei, S. Ananthanarayanan, R. Anubhai et al., "Deep speech 2: end-to-end speech recognition in English and Mandarin," in *Proceedings of the International Conference on Machine Learning*, pp. 173–182, New York, NY, USA, June 2016.
- [2] S. Ren, K. He, R. Girshick et al., "Faster r-cnn: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.
- [3] K. He, X. Zhang, S. Ren et al., "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, Las Vegas, NV, USA, June 2016.
- [4] K. Lee, K. Lee, K. Min et al., "Hierarchical novelty detection for visual object recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1034–1042, Salt Lake City, UT, USA, June 2018.
- [5] I. Evtimov, K. Eykholt, E. Fernandes et al., "Robust physical-world attacks on machine learning models," vol. 2, no. 3, p. 4, 2017, <https://arxiv.org/abs/1707.08945>.

- [6] M. Sharif, S. Bhagavatula, L. Bauer et al., “Accessorize to a crime: real and stealthy attacks on state-of-the-art face recognition,” in *Proceedings of the the 2016 ACM SIGSAC Conference*, Vienna, Austria, October 2016.
- [7] A. Madry, A. Makelov, L. Schmidt et al., “Towards deep learning models resistant to adversarial attacks,” 2017, <https://arxiv.org/abs/1706.06083>.
- [8] A. Lomuscio and L. Maganti, “An approach to reachability analysis for feed-forward relu neural networks,” 2017, <https://arxiv.org/abs/1706.07351>.
- [9] K. Y. Xiao, T. Vincent, N. M. Shafiqullah et al., “Training for faster adversarial robustness verification via inducing relu stability,” 2018, <https://arxiv.org/abs/1809.03008>.
- [10] N. Carlini, G. Katz, C. Barrett et al., “Provably minimally-distorted adversarial examples,” 2017, <https://arxiv.org/abs/1709.10207>.
- [11] K. Lee, K. Lee, H. Lee et al., “A simple unified framework for detecting out-of-distribution samples and adversarial attacks,” *Advances in Neural Information Processing Systems*, vol. 1, pp. 7167–7177, 2018.
- [12] X. Ma, B. Li, Y. Wang, S. M. Erfani et al., “Characterizing adversarial subspaces using local intrinsic dimensionality,” 2018, <https://arxiv.org/abs/1801.02613>.
- [13] W. Xu, D. Evans, and Y. Qi, “Feature squeezing: detecting adversarial examples in deep neural networks,” 2017, <https://arxiv.org/abs/1704.01155>.
- [14] P. Yang, J. Chen, C.-J. Hsieh et al., “ML-LOO: Detecting adversarial examples with feature attribution,” 2019, <https://arxiv.org/abs/1906.03499>.
- [15] K. Roth, Y. Kilcher, and T. Hofmann, “The odds are odd: a statistical test for detecting adversarial examples,” 2019, <https://arxiv.org/abs/1902.04818>.
- [16] N. Papernot and P. McDaniel, “Deep k-nearest neighbors: towards confident, interpretable and robust deep learning,” 2018, <https://arxiv.org/abs/1803.04765>.
- [17] J. J. Engelsma and A. K. Jain, “Generalizing fingerprint spoof detector: learning a one-class classifier,” in *Proceedings of the 2019 International Conference on Biometrics (ICB)*, pp. 1–8, Houston, TX, USA, June 2019.
- [18] K. Grosse, P. Manoharan, N. Papernot et al., “On the (statistical) detection of adversarial examples,” 2017, <https://arxiv.org/abs/1702.06280>.
- [19] I. Goodfellow, J. Pouget-Abadie, M. Mirza et al., “Generative adversarial nets,” *Advances in Neural Information Processing Systems*, vol. 27, pp. 2672–2680, 2014.
- [20] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” 2014, <https://arxiv.org/abs/1412.6572>.
- [21] N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” in *Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57, IEEE, San Jose, CA, USA, May 2017.
- [22] S.-M. Moosavi-Dezfooli and A. Fawzi, “Deepfool: a simple and accurate method to fool deep neural networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2574–2582, Las Vegas, NV, USA, June 2016.
- [23] A. Kurakin, I. Goodfellow, and S. Bengio, “Adversarial samples in the physical world,” 2016, <https://arxiv.org/abs/1607.02533>.
- [24] U. Shaham, Y. Yamada, and S. Negahban, “Understanding adversarial training: increasing local stability of neural nets through robust optimization,” 2015, <https://arxiv.org/abs/1511.05432>.
- [25] C. Szegedy, W. Zaremba, I. Sutskever et al., “Intriguing properties of neural networks,” 2013, <https://arxiv.org/abs/1312.6199>.
- [26] N. Papernot, P. McDaniel, X. Wu et al., “Distillation as a defense to adversarial perturbations against deep neural networks,” in *Proceedings of the 2016 IEEE Symposium on Security and Privacy (SP)*, pp. 582–597, IEEE, San Jose, CA, USA, May 2016.
- [27] M. Jan Hendrik, G. Tim, V. Fischer et al., “On detecting adversarial perturbations,” 2017, <https://arxiv.org/abs/1702.04267>.
- [28] S. Akcay, A. Atapour-Abarghouei, and T. PBreckon, “Ganomaly: Semi-supervised anomaly detection via adversarial training,” in *Asian Conference on Computer Vision*, pp. 622–637, Springer, Perth, Australia, January 2018.
- [29] T. Salimans, I. Goodfellow, W. Zaremba et al., “Improved techniques for training gans,” 2016, <https://arxiv.org/abs/1606.03498>.
- [30] P. Isola, J.-Y. Zhu, T. Zhou et al., “Image-to-image translation with conditional adversarial networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1125–1134, Honolulu, HI, USA, July 2017.
- [31] A. Krizhevsky, V. Nair, and Hinton, “Cifar-10 (canadian institute for advanced research),” <http://www.cs.toronto.edu/kriz/cifar.html>.
- [32] Y. Netzer, T. Wang, A. Coates et al., “Reading digits in natural images with unsupervised feature learning,” 2011, http://ufldl.stanford.edu/housenumbers/nips2011_housenumbers.pdf.
- [33] G. Huang, Z. Liu, L. Van Der Maaten et al., “Densely connected convolutional networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708, Honolulu, HI, USA, July 2017.
- [34] T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii, “Virtual adversarial training: a regularization method for supervised and semi-supervised learning,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1979–1993, 2018.
- [35] G. Xu, Z. Liu, X. Li, and C. Change Loy, “Knowledge distillation meets self-supervision,” in *Proceedings of the European Conference on Computer Vision*, Springer, Cham, Switzerland, 2020.
- [36] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “Mixup: beyond empirical risk minimization,” 2017, <https://arxiv.org/abs/1710.09412>.
- [37] T. Bai, J. Luo, J. Zhao, B. Wen, and Q. Wang, “Recent advances in adversarial training for adversarial robustness,” 2021, <https://arxiv.org/abs/2102.01356>.

Research Article

Cooperative Target Search of UAV Swarm with Communication Distance Constraint

Ning Wang ^{1,2}, Zhe Li ^{1,2}, Xiaolong Liang^{1,2}, Ying Li,³ and Feihu Zhao^{1,2}

¹Air Traffic Control and Navigation College, Air Force Engineering University, Xi'an, China

²Shaanxi Key Laboratory of Electronic Information System Integration, Xi'an 710072, China

³School of Computer Science & Technology, Beijing Institute of Technology, Beijing, China

Correspondence should be addressed to Zhe Li; kongyanshi@126.com

Received 25 June 2021; Accepted 12 August 2021; Published 13 September 2021

Academic Editor: Jie Chen

Copyright © 2021 Ning Wang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This paper proposes a cooperative search algorithm to enable swarms of unmanned aerial vehicles (UAVs) to capture moving targets. It is based on prior information and target probability constrained by inter-UAV distance for safety and communication. First, a rasterized environmental cognitive map is created to characterize the task area. Second, based on Bayesian theory, the posterior probability of a target's existence is updated using UAV detection information. Third, the predicted probability distribution of the dynamic time-sensitive target is obtained by calculating the target transition probability. Fourth, a customized information interaction mechanism switches the interaction strategy and content according to the communication distance to produce cooperative decision-making in the UAV swarm. Finally, rolling-time domain optimization generates interactive information, so interactive behavior and autonomous decision-making among the swarm members are realized. Simulation results showed that the proposed algorithm can effectively complete a cooperative moving-target search when constrained by communication distance yet still cooperate effectively in unexpected situations such as a fire.

1. Introduction

In combat, search and reconnaissance are important for providing effective information to accelerate the observe-orient-decide-act (OODA) cycle [1–4]. Consequently, the US military has identified wide-area target search capability as one of the medium and long-term development goals of unmanned aerial vehicles (UAVs) [3, 4]. UAV swarms have excellent wide-area search capabilities affected by cooperation among swarm members; that is, the collective capability is far greater than the sum of all single UAVs [5, 6]. Cooperative search planning is integral for guiding swarms to achieve wide-area search and target acquisition and has been widely studied [7]. To ensure cooperative search efficiency, a reasonable search planning area and an efficient cooperative strategy are needed.

To determine the area, grid [8, 9], landmark [10, 11], and potential field [12, 13] methods are the main ones proposed. In [8], based on rasterizing the task area, real-time path

planning was realized through an improved ant colony algorithm. In [11], the task area was divided by a Voronoi diagram, and waypoint allocation and track smoothing were used to realize the fast planning of a search track in a static environment. In [13], based on describing the task area using an artificial potential field, an improved logarithmic linear learning algorithm was proposed to reduce the risk that a UAV may wander into a zero-potential field area.

Valente et al. [14] proposed a cooperative search method based on a diffusion-weighted uncertainty model. Each UAV is assigned a search area, and then a potential field algorithm based on a rolling-time domain program solves each search track, but this method can only search for a single moving target. Zhang et al. [15] initialized the target probability distribution map using prior target information and then introduced the environmental uncertainty map to guide the UAV to return to a grid that had not been searched for a long time. The result was a feasible scheme for long-time swarm searches and surveillance track planning. However, only one

kind of prior information, including the initial position of the target, was considered, so the speed and direction of motion were not considered. Dong et al. [16] defined a digital pheromone map and a corresponding updating strategy to realize UAV cooperation in a moving-target search. In [17], a Markov chain was used to describe the target, but it could not consider communication distance and other constraints.

To sum up, the current research enables UAV swarms to have certain cooperative search capabilities, but there are still problems:

- (1) The influence of communication distance on a UAV swarm's cooperation is not considered
- (2) The use of a variety of prior information in a moving-target search is not considered
- (3) The risk of collision in a UAV swarm is simplified as a function of height layers, or it is not considered

In view of the preceding, this paper has done the following work:

- (1) It designed a cooperative search method suitable for dynamic communication distance
- (2) It analyzed the prior information of four typical moving targets to generate a mathematical model that defines them in a cooperative search
- (3) It provided an interface to apply to current UAV conflict resolution results [18–21] in swarm cooperative search missions

2. Description of the Cooperative Moving-Target Search Problem

2.1. Task Description. UAV swarm cooperation is usually divided into area-coverage and target search tasks [22, 23]. The former is to make the UAV swarm complete a flight over a maximum coverage area as soon as possible when prior information about the mission area is difficult to obtain. The latter is to obtain all target information when some prior information, such as location and quantity of targets, is known [24]. This paper is concerned with the second kind of task. In practice, some prior information of enemy target distribution can be obtained from satellite remote sensing imagery and radar detection, which also provides advantages for target acquisition.

Figure 1 is a typical task scenario for a cooperative UAV target search. There are N_t potential moving targets in the mission area, and N_u UAVs are used to search the mission area. Moving targets, such as enemy missile launch vehicles and radar vehicles, are deployed at corresponding positions to protect key enemy targets. Our four UAVs set out from different positions to inspect and defeat moving enemy targets.

2.2. UAV Motion Model. If the UAV swarm that performs the task is U_s , then its motion is given by

$$U_s = \{U_i \mid i = 1, 2, \dots, N_u\}, \quad (1)$$

where i is the number of the swarm members and N_u is the scale of the swarm.

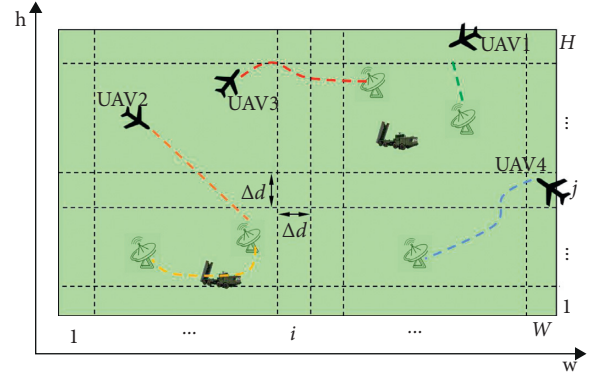


FIGURE 1: Schematic diagram of cooperative UAV swarm search.

To simplify the search decision, a UAV is regarded as a particle in space. The task area is divided into a $W \times H$ grid map where the two-dimensional coordinates (w, h) are used to discretize the UAV motion range and decision set [25]. Assuming that the UAV moves in the grid every time, the constraint of normal overload of UAV movement is satisfied by limiting the grid size and the maximum turning angle, thus ensuring flyability along the planned track. Therefore, the UAV can fly in eight directions at any time, as shown in Figure 2.

The state vector $s_i(k)$ of U_i at k moment is

$$s_i(k) = \begin{bmatrix} x_i(k) \\ y_i(k) \\ \psi_i(k) \end{bmatrix}, \quad (2)$$

where $(x_i(k), y_i(k))$ is the position of U_i at k moment in the environmental awareness map, and $\psi_i(k)$ is the flight course of U_i at k moment, and

$$\psi_i(k) \in \{0, 1, 2, 3, 4, 5, 6, 7\}. \quad (3)$$

Then the flight direction of U_i at k moment can be given by

$$\psi_i(k+1) = (\psi_i(k) + u_i(k)) \bmod 8, \quad (4)$$

where $u_i(k)$ values are the maximum turning angle constraints of the UAV. The state transition function of UAV is then

$$s_i(k+1) = f_1(s_i(k), u_i(k)), \quad (5)$$

where $f_1(\cdot)$ is the UAV state transition function determined by Figure 2 and equation (4).

2.3. Autonomous Decision Function. When performing a cooperative search task amid strong electromagnetic interference, centralized decision-making can realize collaboration among swarm members, but it depends on strict communication, which is often difficult to apply [24]. In this paper, the distributed decision-making method is adopted to make full use of the limited UAV communication distance so that swarm members can make interactive decisions to improve the search and avoid collisions.

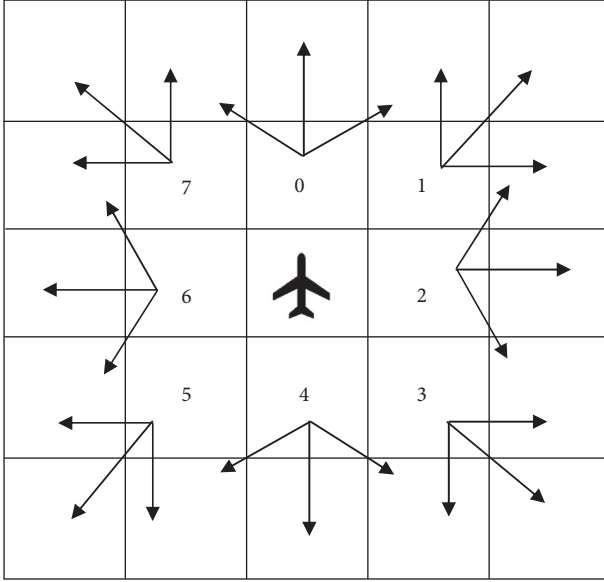


FIGURE 2: UAV flight decision set.

Rolling-time domain optimization decision-making allows a UAV to predict a subsequent multistep decision according to its current state vector and environmental cognitive map at a given moment [26, 27]. The first step prediction is the actual decision quantity, which avoids a decline in overall search efficiency caused by a focus on only short-term search benefits.

If U_i cannot communicate effectively with other members of the swarm at k moment, it uses the differential evolution algorithm to solve the cumulative autonomous decision function according to its current state vector and environmental cognitive map, thereby obtaining the current decision value. The autonomous decision function of U_i at k moment is

$$J_{\text{ind}}^i(k) = w_1 J_1 + w_2 J_2 + w_3 J_3, \quad (6)$$

s.t. $k \in [0, T_a]$; $i = 1, 2, \dots, N_u$,

where J_{ind}^i is composed of target search revenue J_1 , environmental search income J_2 , and expect probe revenue J_3 ; w_l indicates the weight of each income in the process of generating interactive information to satisfy $w_l \in [0, 1]$ and $\sum_{l=1}^3 w_l = 1, l = 1, 2, 3$, and values according to the specific task requirements and engineering experience. The benefits of autonomous decision function are described as follows.

2.3.1. Probability Return of Target Existence J_1 . Target existence probability income represents the value of possible targets in the corresponding environmental cognitive map, which guides the UAV to search areas that have a high possibility of targets. It is defined as

$$J_1(k) = \sum_{w=1}^W \sum_{h=1}^H (p_{wh}(k) - p_{wh}(k-1)), \quad (7)$$

where $p_{wh}(k)$ indicates the probability of having a goal in grid (w, h) at k moment that satisfies $p_{wh} \in [0, 1]$.

2.3.2. Income from Environmental Uncertainty J_2 . Environmental uncertainty revenue represents the reduction of uncertainty in the grid of the corresponding environmental cognitive map after the UAV has searched it. It guides the UAV to search the task area with high uncertainty and reduces the possibility of missing targets:

$$J_2(k) = \sum_{w=1}^W \sum_{h=1}^H (\psi_{wh}(k+1) - \psi_{wh}(k)), \quad (8)$$

where $\psi_{wh}(k)$ represents the uncertainty of grid (w, h) in the environment cognitive map, and $\psi_{wh}(k) \in [0, 1]$ (see Section 3.2 for details).

2.3.3. Comprehensive Income J_3 . Comprehensive income is obtained by multiplying environmental uncertainty and target existence probability, which is used to guide the UAV to detect areas with high uncertainty and target existence probability. The detailed description is [15]

$$J_3(k) = \sum_{w=1}^W \sum_{h=1}^H (p_{wh}(k+1) \cdot \psi_{wh}(k+1) - p_{wh}(k) \cdot \psi_{wh}(k)). \quad (9)$$

Then, the autonomous decision value $u_i^*(k)$ of U_i at k moment can be given by the following formula:

$$u_i^*(k) = \arg \max \sum_{t=k}^{t=k+q-1} J_{\text{ind}}^i(s_i(t), E^i(t)), \quad (10)$$

where $s_i(t)$ is the state vector of U_i at t moment, $E^i(t)$ is the environmental cognitive map of U_i at t moment, and q is the step size of the rolling-time domain.

2.4. Interactive Decision Function. When U_i is in the communication range of other swarm members, the efficiency of cooperative search can be improved by sharing information, but the problem of collision prevention should be considered. The flight conflict resolution method and information interaction method are described below.

There has been much research on collisions in UAV swarms in the distributed decision framework. The literature [12] guides UAVs to avoid conflict by establishing an artificial potential field, which has the characteristic of a short response time and requires only a small amount of calculation. It can realize real-time obstacle avoidance but cannot resolve complex conflict problems. In [18–20], the speed obstacle method broadcasts automatic correlation monitoring to give each UAV the position and speed of the others; thus it solved the potential problem by detecting flight conflict and determining a relief flight path; however, the relief path can easily deviate the UAV from the search target point, thus compromising mission efficiency. In [21], the distributed model predictive control method is adopted, in which the collision avoidance management unit and the interactive graph updating mechanism address conflict resolution in multi-UAV route planning, but it requires a large amount of computation.

To study cooperative strategy in the search for moving targets, this paper introduced the artificial potential field term to meet the basic requirements of collision prevention. In practice, the minimum safe distance can be defined by this method. To achieve a better anticollision effect, the previous conflict resolution decision method replaced the interactive decision function within the minimum safe distance.

The anticollision constraint J_4 is defined as

$$J_4 = \lg \left(\sum_{j=1}^{j=N_u} \frac{l_{ij}}{N_{\text{comu}}} \right), \quad (11)$$

where N_{comu} represents the number of members that can communicate with U_i (see Section 3.2) and l_{ij} indicates the distance between U_i and the other swarm members within the communication distance of U_i , which is given by

$$l_{ij} = \begin{cases} 0, & c_{ij}(k) = 0, \\ \|s_i(k) - s_j(k)\|, & c_{ij}(k) = 1, \end{cases} \quad (12)$$

where $c_{ij} = 0$ means that U_j is not within the direct communication distance of U_i at k moment; $c_{ij} = 1$ means that U_j can communicate directly with U_i ; $\|\cdot\|$ is the second norm, which is used to calculate the distance between two UAVs that can communicate directly. Furthermore, it is possible to obtain the interactive decision function of U_i as

$$J_{\text{int}}^i(k) = w_1 \cdot J_1 + w_2 \cdot J_2 + w_3 \cdot J_3 + w_4 \cdot J_4, \quad (13)$$

where J_{int}^i is composed of target search revenue J_1 , environmental search income J_2 , expected detection income J_3 , and anticollision constraint J_4 ; w_l indicates the weight of each income in the process of generating interactive information that satisfies $w_l \in [0, 1]$ and $\sum_{l=1}^3 w_l = 1, l = 1, 2, 3, 4$.

Then the interactive decision value $u_i^*(k)$ of U_i can be given by the following formula:

$$u_i^*(k) = \arg \max \sum_{t=k}^{t=k+q-1} J_{\text{int}}^i(s_i(k), E_{\text{inter}}^i(k)), \quad (14)$$

where $s_i(k)$ is the state vector of U_i at k moment, $E_{\text{inter}}^i(k)$ is a decision-making environment cognitive map fused to the environment map of other members acquired by U_i at k moment (see Section 4.1), and q is the optimized step size of the rolling-time domain.

3. Construction and Update of Environmental Cognition Map

In a UAV swarm search, environmental cognitive maps (target probability distribution and environmental

uncertainty maps) are used to describe the environmental state, and swarm members interact with each other through their own environmental cognitive maps.

3.1. Target Probability Distribution Map Initialization and Update. In a cooperative search, the existence probability of the target in grid (w, h) at k moment can be expressed as $p_{wh}(k) \in [0, 1]$. Among them, $p_{wh}(k) = 0$ expresses no target in grid (w, h) at k moment, whereas $p_{wh}(k) = 1$ expresses targets in grid (w, h) at k moment. Now, the target probability distribution diagram of U_i at k moment can be expressed as

$$P_i(k) = \{p_{wh}^i(k) \mid w = 1, 2, \dots, W, h = 1, 2, \dots, H\}. \quad (15)$$

In order to make full use of the prior information of the moving target, we divide the target in the cooperative search task into four types, as shown in Table 1.

3.1.1. Unknown Target Position and Speed Information. At this time, the probability distribution of targets in the task area is uniform, and the probability density function of any target in the task area can be expressed as

$$f(x, y) = \frac{1}{(W \cdot H)}. \quad (16)$$

3.1.2. The Initial Position of the Target Is Known, but the Velocity Information Is Unknown. Assume that the task area has N_2 targets that have type 2 prior information, and $(x_{\text{tar}}^{n2}, y_{\text{tar}}^{n2})$ is used to represent its initial position. Each such target can be considered to obey a two-dimensional normal distribution $N(x_{\text{tar}}^{n2}, y_{\text{tar}}^{n2}, \delta_0^2, \delta_0^2, \rho)$, because the distribution of x, y is independent of each other and $\rho = 0$. Without losing generality, assuming that the initial position distribution of each target is independent, the total probability distribution density of the second target can be expressed as

$$f(x, y) = \sum_{n2=1}^{N_2} \frac{1}{2\pi \delta_0^2} \cdot \exp^{-((x-x_{\text{tar}}^{n2})^2/2\delta_0^2 + (y-y_{\text{tar}}^{n2})^2/2\delta_0^2)}. \quad (17)$$

After t_0 time, the target moves from the initial position in an independent, incremental process [16] as described by the Wiener stochastic process: $x_{\text{tar}}^{n2}(t) \sim N(0, \delta_e^2 t_0)$, $y_{\text{tar}}^{n2}(t) \sim N(0, \delta_e^2 t_0)$. The distribution density probability of the second kind of target at t_0 time is

$$f(x, y) = \sum_{n2=1}^{N_2} \frac{1}{(2\pi \delta_0^2 + \delta_e^2 t_0)} \cdot \exp^{-((x-x_{\text{tar}}^{n2})^2/2(\delta_0^2 + \delta_e^2 t_0) + (y-y_{\text{tar}}^{n2})^2/2(\delta_0^2 + \delta_e^2 t_0))}. \quad (18)$$

TABLE 1: Classification of prior information of targets.

	Target initial position	Target speed size	Target speed direction
Type 1	×	×	×
Type 2	√	×	×
Type 3	√	√	×
Type 4	√	√	√

3.1.3. *The Initial Position and Velocity of the Target Are Known, but the Moving Direction Is Unknown.* Assume that the task area has N_3 targets having type 3 prior information, and each initial position is expressed by $(x_{\text{tar}}^{n_3}, y_{\text{tar}}^{n_3})$ and speed

$$f(x, y) = \sum_{n_3=1}^{N_3} \frac{1}{(2\pi\delta_0)^2} \int_{\theta=0}^{2\pi} \exp^{-((x+v_{n_3}t_0 \cos \theta - x_{\text{tar}}^{n_3})^2/2\delta_0^2 + (y+v_{n_3}t_0 \sin \theta - y_{\text{tar}}^{n_3})^2/2\delta_0^2)} d\theta. \quad (20)$$

3.1.4. *The Initial Position, Velocity Direction, and Size of the Target Are Known.* Assume that the task area has N_4 targets with type 4 prior information. The initial position is $(x_{\text{tar}}^{n_4}, y_{\text{tar}}^{n_4})$, speed size is v_{n_4} , and $\theta_{n_4} \in [0, 2\pi]$ represents the

$$f(x, y) = \sum_{n_4=1}^{N_4} \frac{1}{2\pi\delta_0^2} \cdot \exp^{-((x+v_{n_4}t_0 \cos(\theta_{n_4}) - x_{\text{tar}}^{n_4})^2/2\delta_0^2 + (y+v_{n_4}t_0 \sin(\theta_{n_4}) - y_{\text{tar}}^{n_4})^2/2\delta_0^2)}. \quad (21)$$

3.2. *Initialization and Update of Environmental Uncertainty Map.* The environmental uncertainty in the grid (w, h) at k moment can be expressed as $\psi_{wh}(k) \in [0, 1]$, where $\psi_{wh}(k) = 1$ represents the information of the grid (w, h) that is completely uncertain at k moment and $\psi_{wh}(k) = 0$ represents the information of the grid (w, h) that is completely obtained by a UAV at k moment. The environmental uncertainty diagram of U_i at k moment can be expressed as

$$\psi_i(k) = \{\psi_{wh}^i(k) \mid w = 1, 2, \dots, W, h = 1, 2, \dots, H\}. \quad (22)$$

The initial environmental uncertainty map is defined as an all-1 matrix. With the increasing number of UAV searches, the grid uncertainty continues to decline. The specific update method is

$$\psi_{wh}(k) = \eta^{\delta n} \cdot \psi_{wh}(k-1), \quad (23)$$

where $\eta \in [0, 1]$ indicates the attenuation factor of environmental uncertainty [28]. It is used to characterize the amount of information obtained by the UAV after searching the corresponding grid once; $\delta n \in \mathbb{N}$ is the number of the grid (w, h) that is searched at k moment.

4. Swarm Cooperation Strategy

Under the distributed cooperative search architecture, when U_i reaches the communication range of other members of

size is represented by v_{n_3} . Since the speed of the target is fixed after t_0 the probability density of the grid (w, h) is transferred from the probability distribution of the arc region of the $v_{n_3}t_0$ radius by (w, h) , namely,

$$f(x, y) = \frac{1}{2\pi v_{n_3} t_0} \cdot \int_L f_0(x_0, y_0) ds, \quad (19)$$

where L is an arc region with center (x_0, y_0) and the radius $v_{n_3}t_0$. By transforming using the first curve integration, the probability distribution of the third-class target area can be obtained as

speed direction. After t_0 , the target position offset is $(v_{n_4}t_0 \cos(\theta_{n_4}), v_{n_4}t_0 \sin(\theta_{n_4}))$, and the probability distribution density of the fourth target can be expressed as

the swarm, it is possible to avoid searching the searched grids repeatedly through information interaction, thus improving the efficiency of the swarm cooperative search. When making interactive decisions under distributed architecture, the decision of a UAV does not depend on the state information of other UAVs or the operation of central nodes. Therefore, this interactive decision-making method can be applied to a strong confrontation environment that has dynamic changes of effective communication distance.

4.1. *Interactive Information Fusion Method.* When performing the cooperative search, the target probability maps of all members can be considered to be updated synchronously because the prior information of the target is completely shared. However, the state of the environmental uncertainty map changes in real time with the search, which requires interaction in the cooperative search. The environmental cognitive map of U_j after realizing information interaction is then defined as

$$E_{\text{inter}}^j(k) = \{\psi_{\text{inter}}^j(k), P_j(k)\}, \quad (24)$$

where the target probability distribution map $P_j(k)$ is updated according to the task execution time and target prior information formulas (16)–(21); $\psi_{\text{inter}}^j(k)$ indicates the interactive environment uncertainty map. If the information

interaction between U_i and U_j is taken as an example, the interaction mode can be given by

$$\psi_{inter}^i(k) = \psi_j(k) \cdot \lceil \psi_i(k) - \psi_j(k) \rceil + \psi_i(k) \cdot \lceil \psi_j(k) - \psi_i(k) \rceil, \quad (25)$$

where $\lceil \cdot \rceil$ is an upward rounding function; $\psi_i(k)$ is the environment uncertain map carried by UAV that can communicate with U_j at k moment.

4.2. Swarm Communication Topology. When UAVs send messages to each other in the form of a swarm broadcast, the distance between U_i and U_j at k moment can be expressed as

$$d_{ij}(k) = \sqrt{|x_i(k) - x_j(k)|^2 + |y_i(k) - y_j(k)|^2}. \quad (26)$$

The finite set of UAVs that can communicate with U_i at k moment can be expressed as

$$C_i(k) = \{U_j(k) \mid c_{ij}(k) = 1\} = \{U_j(k) \mid U_j \in U_s \cap d_{ij}(k) \leq R(k)\}, \quad (27)$$

where $c_{ij}(k)$ represents the communication state between U_i and U_j and $R(k)$ is the effective communication distance among swarm members at k moment.

4.3. Algorithm Flow. The pseudocode of a UAV swarm moving-target search algorithm with communication distance constraint is shown in Algorithm 1, and the specific steps are described as follows:

Step 1: environmental awareness map and parameter initialization. The environmental cognitive map is initialized according to the prior information of four kinds of moving objects. The rolling-time domain optimization step of the UAV is q ; the initial effective communication distance is $R(0)$; the scale of the UAV swarm is N_u ; and the initial state, course, and weight coefficient of the autonomous decision and interactive decision functions of the UAV are set.

Step 2: autonomous decision. According to the UAV's own independent decision function and environmental cognitive map, the differential evolution algorithm is used to solve the problem and make real-time route planning.

Step 3: information interaction. When the UAV reaches the communication range of the other members of the swarm, the interactive information of the other members is fused by formula (25), and the cognitive map of its own decision environment is updated.

Step 4: interactive decision. After updating the self-knowledge map of the environment according to the interactive decision function in formula (11), the differential evolution algorithm is used to track in real time.

Step 5: update the target probability map. In the cooperative search process of a UAV swarm, according to UAV decision information and the target probability

map updating method, the target probability distribution map is updated by formula (16) to formula (21), and the environment uncertain map is updated by formula (23).

Step 6: repeat step 2 to make the next decision based on the updated environmental cognitive map.

5. Comparative Analysis of Simulation

In this section, concerning the moving-target search scene with four types of prior information, the UAV swarm cooperative search was simulated numerically, with the impact of introducing prior target information. In the cooperative search, the task was simulated and analyzed, and the effectiveness of the algorithm in the strong confrontation environment was verified, such as the dynamic change of the communication distance, the damage of some members of the swarm, and other emergencies.

5.1. Task Assumption and Parameter Setting. The reconnaissance mission area is a 30×40 km rectangle divided into 1×1 km grids. The initial distribution, speed direction, and performance constraints of a UAV swarm are shown in Table 2, and the parameters of autonomous decision-making and interactive decision-making are shown in Table 3 and Table 4, respectively. Set the simulation time to 6000 s and the rolling-time domain optimization step to 30 s. The simulation time is divided into 600 planning steps with an interval of 10 s. The initial effective UAV communication distance is set at 3 km. According to the predetermined prior information, the initial target location distribution is shown in Figure 3, the speed of the target is 10 km/h, and the target probability distribution generated by the UAV swarm according to the prior information is shown in Figure 4.

5.2. Planning Results of Moving-Target Cooperative Search. Given a limited communication distance, the numerical simulation of a cooperative moving-target search is carried out using the number of captured targets as an evaluation index. The simulation results are shown in Figure 5.

Figure 5(a) shows that the swarm captured two type 4 targets after 1000 s based on prior information and another target had moved out of the task area, so the target probability distribution is concentrated. It attracted the attention of the UAV swarm and was then captured. In Figure 5(b), after the swarm captured the category 4 targets, it quickly carried out a cooperative search in the center of the task area where other target categories were concentrated. According to Figures 5(c) and 5(d), when the task was executed at 6000 s, the swarm completed coverage of the task area and captured 1 type-1 target, 1 type-2 target, 2 type-3 targets, and 3 type-4 targets. Because prior information of the type-1 moving target was unknown, it was difficult to capture, but the swarm captured other moving targets by making full use of prior information: the richer the prior information, the higher the capture probability.

```

main program
(1) Initialize algorithm parameters, environment map, UAV position, and heading
(2) for tar = 1:  $N_{tar}$ 
(3) if tar in class 1
(4) Initialize the environmental cognitive map according to formula (16);
(5) ...
(6) if tar in class 4
(7) Initialize the environmental cognitive map according to formula (21);
(8) end if
(9) end for
(10) for  $k = 0: k_{max}$ 
(11) Update the target probability distribution map according to equations (16)–(21);
(12) for  $U_i = 1: N_u$ 
(13) According to formula (26), judge whether to make an independent decision;
(14) According to formula (11), a differential evolution algorithm is adopted to make a real-time decision;
(15) for  $U_j = 1: N_u$ 
(16) Determine the interactive member set according to formula (27);
(17) Complete information interaction and fusion according to formula (25);
(18) end for
(19) According to formula (14), the differential evolution algorithm is adopted to make real-time decisions;
(20) Update the UAV position according to formula (5);
(21) Update your own environmental cognition map according to formula (23);
(22) end for
(23) end for

```

ALGORITHM 1: Algorithm pseudocode.

5.3. Search Path Planning under Dynamic Communication Conditions. Electromagnetic interference (EMI) is a regular means of attack in a strong confrontation environment, and it has a severe influence on battlefield communication. To verify the applicability of the algorithm in this paper to a complex communication environment, this section simulated and analyzed the cooperative search of a UAV swarm under a dynamically changing effective communication distance. The total simulation time (6000 s) was divided into 600 planning steps at an interval of 10 s. The effective initial intermachine communication distance was 20 km. After 1500 s, the distance was reduced to 10 km and restored to 2000 s. The effective communication distance was reduced to 0 km when the task was executed at 3500 s and was restored at 4500 s.

To further demonstrate the applicability of the algorithm in a dynamic communication environment, this section took environmental uncertainty as the main goal guiding the swarm and used the search coverage rate of the task area as the evaluation index to carry out a numerical simulation. At this time, the algorithm simulation parameters were updated (Tables 5 and 6).

The left half of Figure 6 shows that when the search went to 1500 s because of decreased effective communication distance, the frequency of information interaction among the UAVs also decreased, but after communication was restored at 2000 s, the number of information interactions was quickly restored. From the right half of Figure 6, when the search went for 3500 s, the distance between computers decreased, which led to slow growth in the area-coverage rate. This was caused by the repeated searching of some grids after interference led to the loss of information interaction.

At 4500 s, communication and interactive decision-making were restored, and search coverage gradually improved, which shows that interactive decision-making can improve cooperative efficiency but that a UAV can still perform scheduled tasks autonomously when communication conditions are not guaranteed.

5.4. Search Route Planning When Some Members of the Swarm Are Damaged. Compared with the preplanning method (referring to the related documents of preplanning), the dynamic planning method can effectively reduce the risk that the enemy can predict and attack the track and adapt to unexpected situations such as the failure of some members. In this section, the parameters of target probability distribution (Tables 3 and 4) and environmental uncertainty (Tables 5 and 6) guide the swarm and take the number of captured targets and the coverage rate as evaluation indexes for the numerical simulation. In this scenario, UAV1 and UAV3 failed at 1500 s and 3500 s, respectively, and stopped executing tasks.

From Figures 7(a)–7(c), when UAV1 and UAV3, respectively, stop performing tasks due to faults, UAV2 and UAV4 still performed effectively and achieved higher regional coverage because in the distributed decision-making architecture each UAV does not depend on other members to make decisions. When the swarm is large, the efficiency of a cooperative search can be improved through interactive decision-making. It can be seen from Figure 7(d) that when some members are damaged, which leads to a decline in swarm size, a UAV can still carry out search tasks through autonomous decision-making, which has good robustness.

TABLE 2: Initial states and performance constraints of UAV.

UAV serial number	Initial coordinates	Initial direction	Fixed flight speed	Maximum turning angle
1	(5, 0)	0°	20	45°
2	(35, 28)	0°	20	45°
3	(5, 29)	180°	20	45°
4	(35, 0)	180°	20	45°

TABLE 3: Independent decision-making parameters.

Parameter	w_1	w_2	w_3	q
Value	0.3	0.3	0.4	3

TABLE 4: Interactive decision parameters.

Parameter	w_1	w_2	w_3	w_4	q
Value	0.2	0.2	0.2	0.4	3

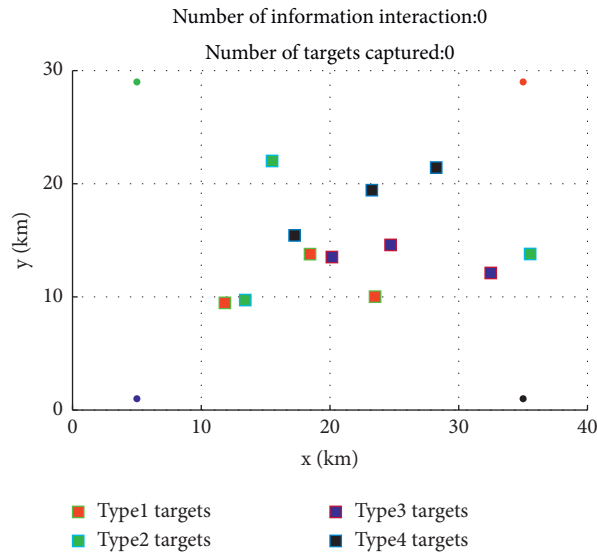


FIGURE 3: Initial position distribution of targets.

Figure 8 shows that, driven by prior information, the UAV swarm searches for and quickly captures type-4 targets according to prior information. It can be seen from the whole search process from Figures 8(a)–8(d), even after UAV1 and UAV3 quit at 1500 s and 3000 s, UAV2 and UAV4 still searched effectively.

5.5. Influence of Communication Distance on Search Efficiency. On the basis of completing the path planning of the cooperative moving-target search using the coverage rate as the main evaluation index, the efficiency under different communication distances was analyzed by using the control variable method. Five groups of simulations having a 10 km

communication interval over 6000 s were carried out and the results are shown in Figure 9.

According to the simulation results, information interaction avoided the repeated search of the same grid, and the cooperative efficiency of the UAV swarm improved. With the increase in communication distance, the cooperative search efficiency gradually increases, but when the communication distance was greater than 30 km the efficiency no longer increased. The results showed a positive correlation in the nonlinear relationship of communication distance to cooperative search efficiency. In practice, the minimum effective communication distance can be preliminarily determined by the simulation to achieve better task cooperation.

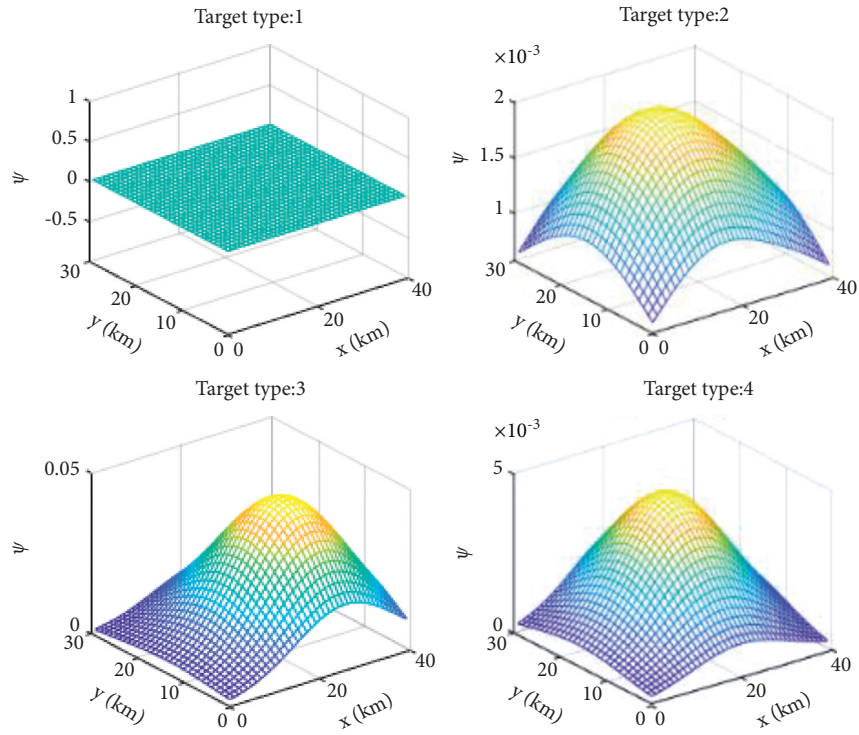


FIGURE 4: Distribution of prior information of initial target.

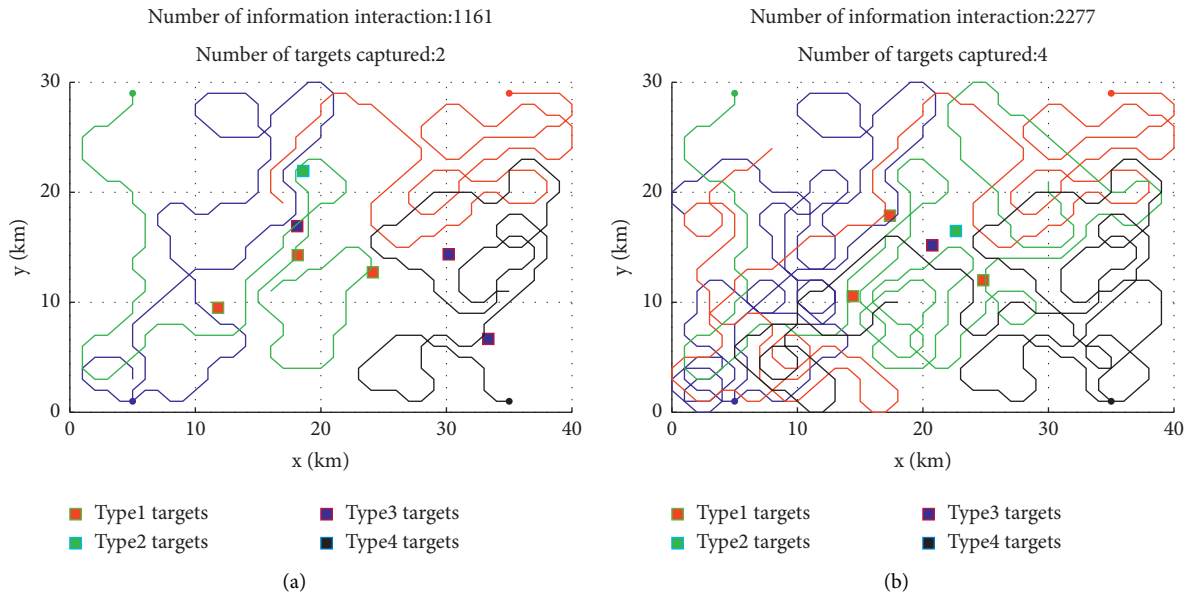


FIGURE 5: Continued.

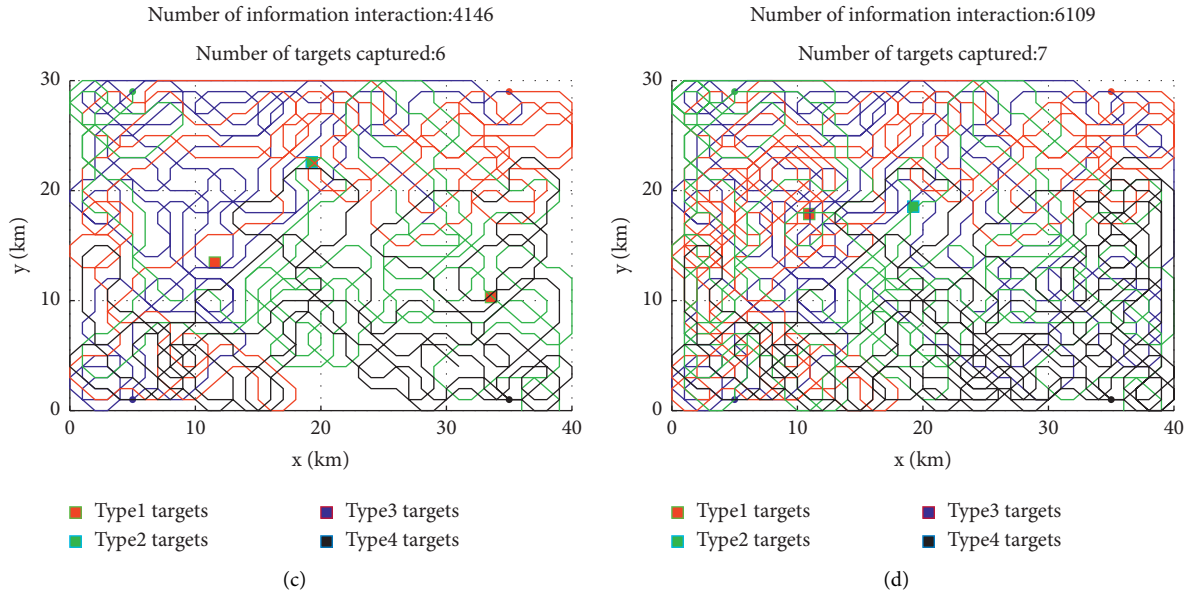


FIGURE 5: Cooperative search planning of moving targets: (a) 1000 s, (b) 2000 s, (c) 4000 s, and (d) 6000 s.

TABLE 5: Independent decision-making parameters.

Parameter	w_1	w_2	w_3	q
Value	0	1	0	3

TABLE 6: Algorithm parameters.

Parameter	w_1	w_2	w_3	w_4	q
Value	0	0.5	0	0.5	3

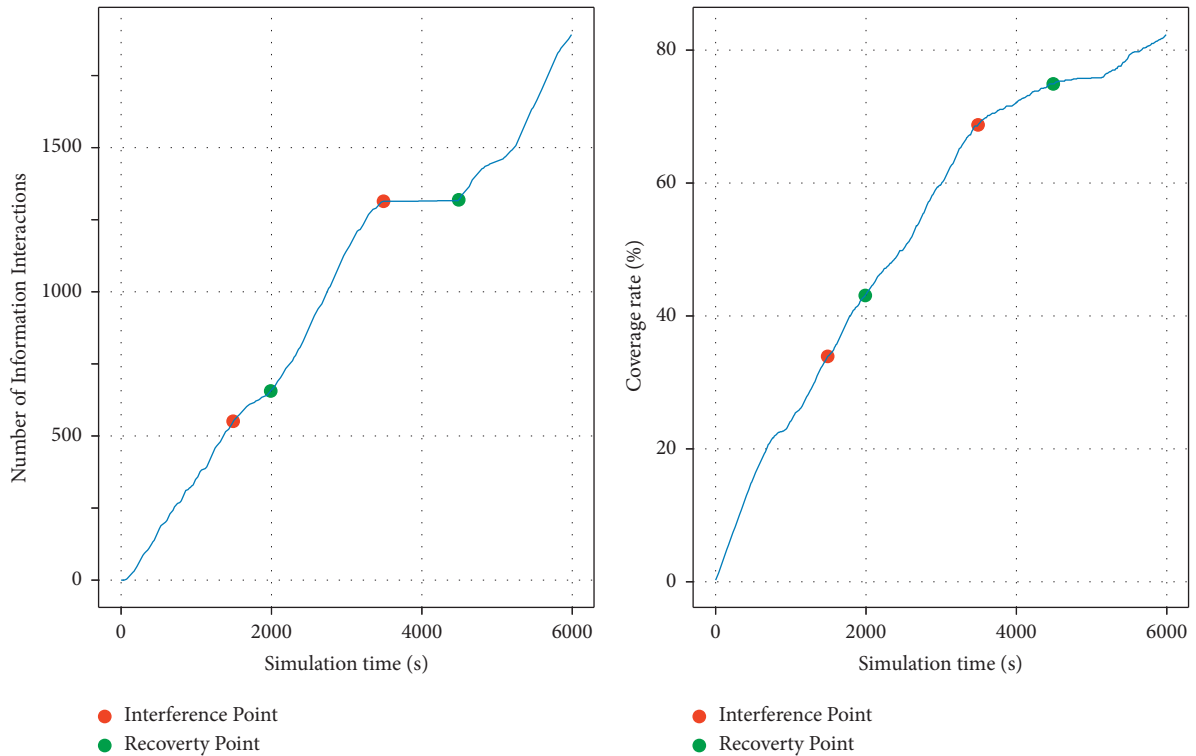


FIGURE 6: Coverage change under communication restriction.

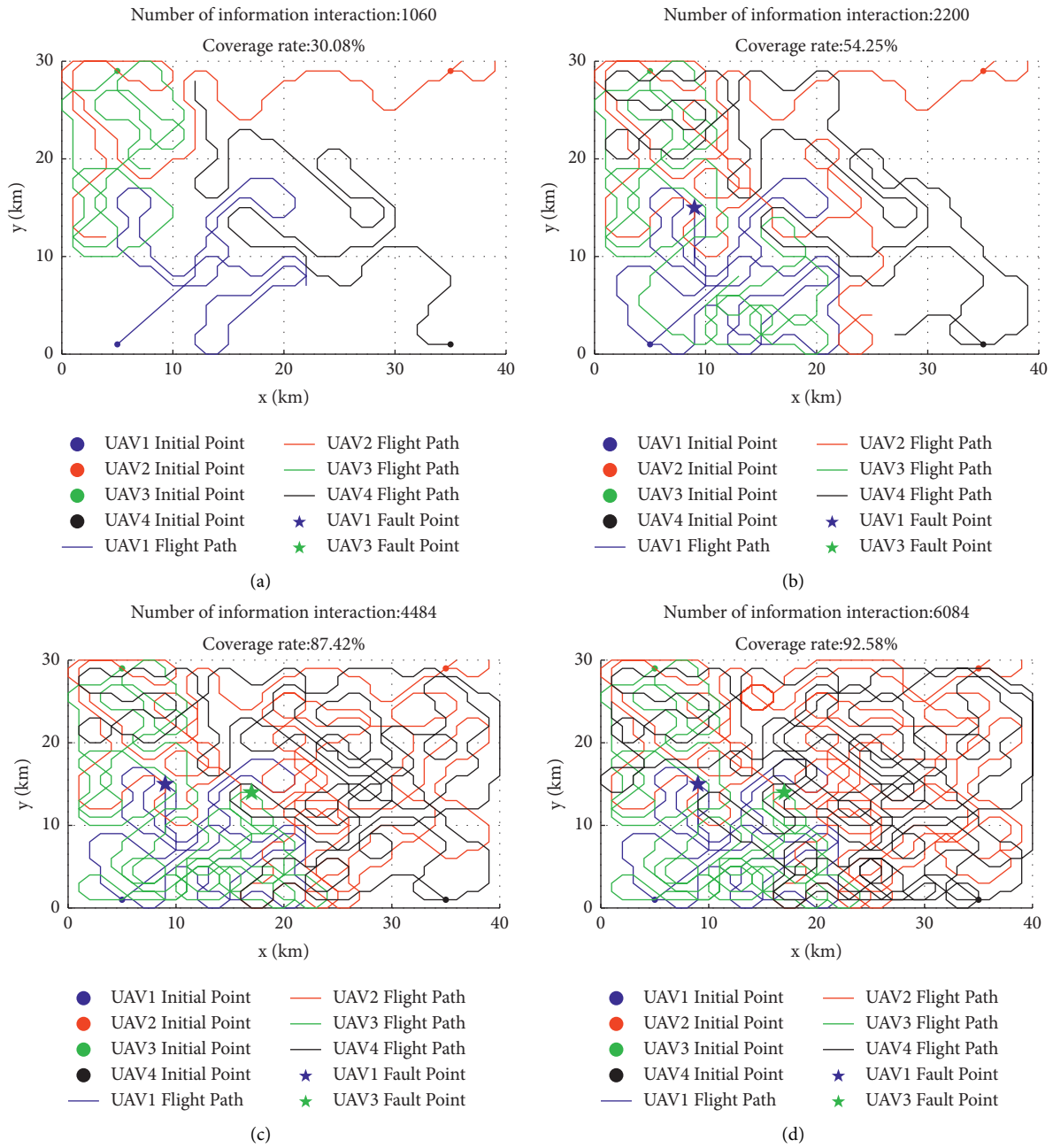


FIGURE 7: Coverage track planning under the condition of partial member damage: (a) 1000 s, (b) 2000 s, (c) 4000 s, and (d) 6000 s.

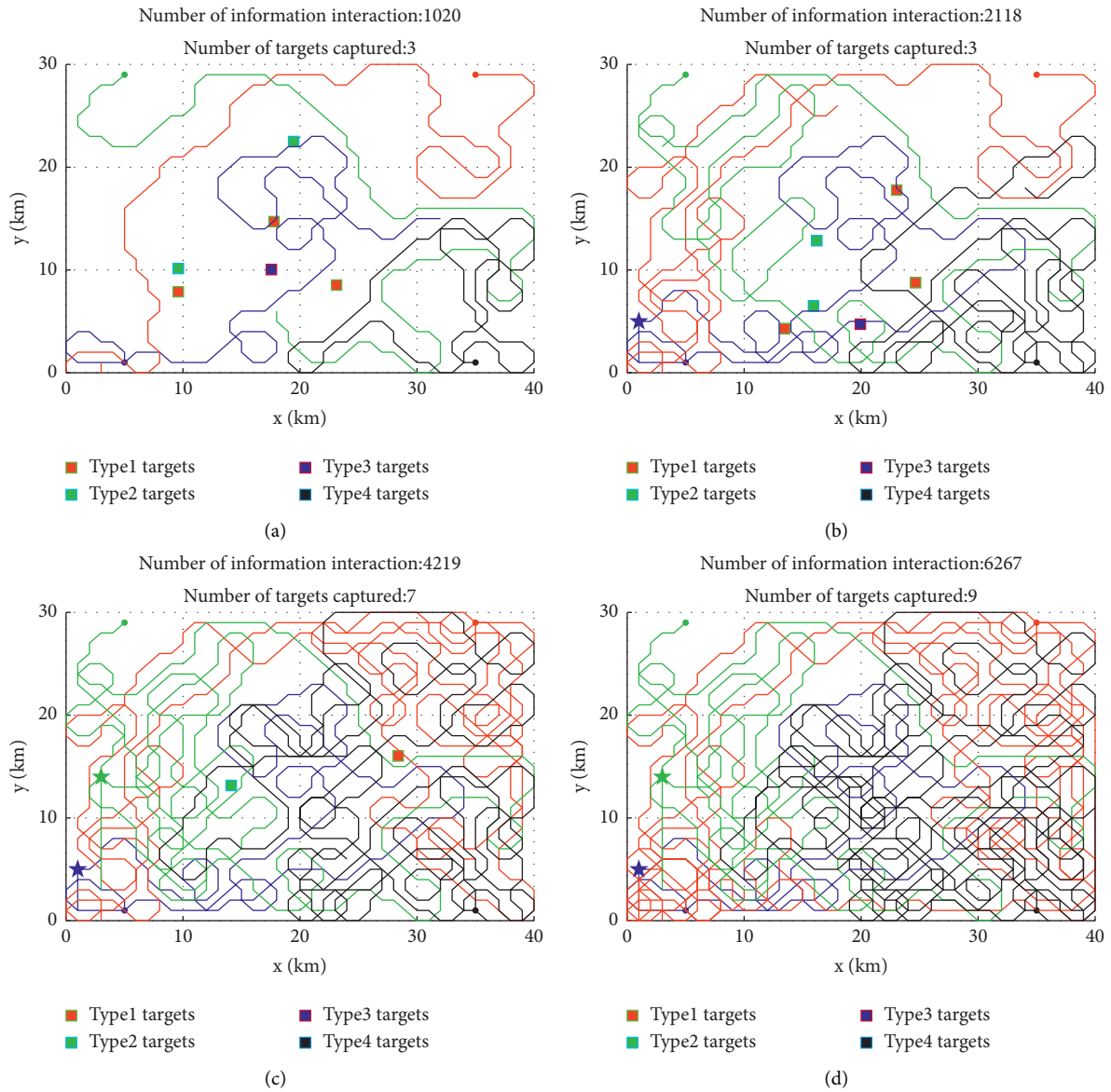


FIGURE 8: Track planning of moving-target search after damage to some members: (a) 1000 s, (b) 2000 s, (c) 4000 s, and (d) 6000 s.

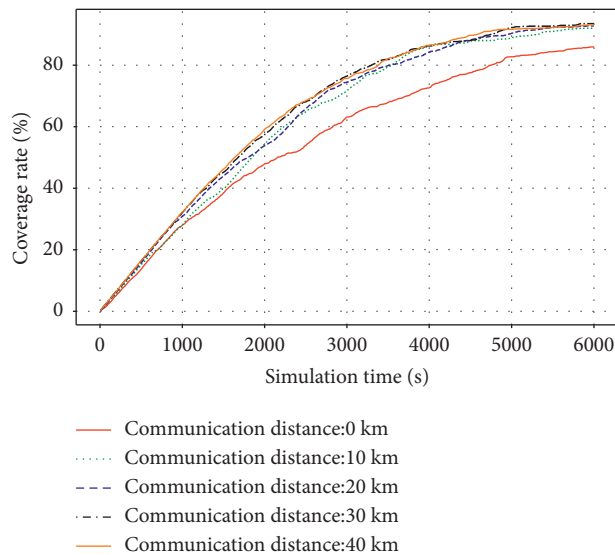


FIGURE 9: Coverage growth under different communication distances.

6. Conclusion

- (1) This algorithm simulated the cooperative search decision-making of a UAV swarm at a limited communication distance and improved its robustness in a strong countermeasure environment by establishing autonomous and interactive decision-making
- (2) The mathematical models and updating methods of prior information of the four types of moving targets were established so that a UAV swarm could make full use of prior information to carry out a cooperative search
- (3) Using a distributed control architecture in an experimental simulation, the algorithm in this paper proved that a cooperative search task can still be completed effectively when some members fail

The algorithm did not consider the influence of communication delay and packet loss on the cooperative search efficiency, but it could in subsequent research to improve the cooperative search algorithm of the UAV swarm.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Authors' Contributions

Ning Wang, Zhe Li, and Feihu Zhao were responsible for conceptualization, methodology, software, and validation; Ying Li and Feihu Zhao performed data curation. Ning Wang, Xiaolong Liang, and Ying Li prepared the original draft and reviewed and edited the manuscript. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

This work was funded by the National Natural Science Foundation of China, Grant no. 61703427.

References

- [1] Office of the Secretary of Defense, *Unmanned Aircraft Systems Roadmap 2005-2030*, Department of Defense, Washington, DC, USA, 2005.
- [2] Office of the Secretary of Defense, *Unmanned Aircraft Systems Roadmap 2007-2032*, Department of Defense, Washington, DC, USA, 2007.
- [3] Office of the Under Secretary of Defense, *Defense Science Board Study on Unmanned Aerial Vehicles and Uninhabited Combat Aerial Vehicles*, Office of the under Secretary of Defense for Acquisition, Technology, and Logistics, Washington, DC, USA, 2004.
- [4] Y. Alshuler, A. Pentland, and M. B. Alfred, *Swarms and Network Intelligence in Search*, Springer International Publishing, Cham, Switzerland, 2018.
- [5] Y. Q. Hou, X. L. Liang, Y. L. He, and J. Q. Zhang, "Time-coordinated control for unmanned aerial vehicle swarm cooperative attack on ground-moving target," *IEEE Access*, vol. 7, pp. 106930–106939, 2019.
- [6] C. C. Cheng, G. H. Bai, Y. A. Zhang, and J. Y. Tao, "Resilience evaluation for UAV swarm performing joint reconnaissance mission," *Chaos*, vol. 5, 2019.
- [7] N. Nigam, S. Bieniawski, I. Kroo, and J. Vian, "Control of multiple UAVs for persistent surveillance: algorithm and flight test results," *IEEE Transactions on Control Systems Technology*, vol. 20, no. 5, pp. 1236–1251, 2012.
- [8] Z. Zhen, Y. Chen, L. Wen, and B. Han, "An intelligent cooperative mission planning scheme of UAV swarm in uncertain dynamic environment," *Aerospace Science and Technology*, vol. 100, Article ID 105826, 2020.
- [9] Z. Lv, L. Yang, Y. He, Z. Liu, and Z. Han, "3D environment modeling with height dimension reduction and path planning for UAV," in *Proceedings of 9th International Conference on Modelling, Identification and Control*, pp. 734–739, Kunming, China, July 2017.
- [10] S. Medeirosfl, "Computational modeling for automatic path planning based on evaluations of the effects of impacts of UAVs on the ground," *Journal of Intelligent & Robotic Systems*, vol. 61, no. 1, pp. 181–202, 2011.
- [11] P. Stodola, J. Drozd, J. Nohel, J. Hodický, and D. Procházka, "Trajectory optimization in a cooperative aerial reconnaissance model," *Sensors*, vol. 19, no. 12, p. 2823, 2019.
- [12] E. C. Lu and W. X. Zhang, "Path planning for mobile robot based on improved artificial potential field method in complex environment," *Computer Engineering and Applications*, vol. 24, pp. 45–48, 2013.
- [13] J. Ni, G. Tang, Z. Mo, W. Cao, and S. X. Yang, "An improved potential game theory based method for multi-UAV cooperative search," *IEEE Access*, vol. 8, pp. 47787–47796, 2020.
- [14] J. Valente, A. Barrientos, J. D. Cerro et al., "Multi-robot visual coverage path planning: geometrical metamorphosis of the workspace through raster graphics based approached," in *Proceedings of International Conference on Computational Science and its Applications (ICCSA)*, pp. 58–73, Santander, Spain, June 2011.
- [15] Z. Zhang, L. Teng, and G. Xu, "Search method for cooperative moving targets of multiple UAVs driven by revisiting mechanism," *Acta Aeronautica Sinica*, vol. 41, no. 5, pp. 220–232, 2020.
- [16] X. Dong, J. Jiang, J. Zhou, and C. Yu, "Target search for cooperative movement of multiple UAVs with limited communication," *Journal of Harbin Engineering University*, vol. 39, no. 11, pp. 1823–1829, 2018.
- [17] X. Fang, S. Ma, Q. Yang, and J. Zhang, "Cooperative energy dispatch for multiple autonomous microgrids with distributed renewable sources and storages," *Energy*, vol. 99, pp. 48–57, 2016.
- [18] Y. Huang, J. Tang, and S. Lao, "Collision avoidance method for self-organizing unmanned aerial vehicle flights," *IEEE Access*, vol. 7, pp. 85536–85547, 2019.
- [19] S. Huang, R. S. H. Teo, and K. K. Tan, "Collision avoidance of multi unmanned aerial vehicles: a review," *Annual Reviews in Control*, vol. 48, pp. 147–164, 2019.
- [20] Y. Huang, J. Tang, and S. Lao, "UAV group formation collision avoidance method based on second-order consensus

- algorithm and improved artificial potential field,” *Symmetry*, vol. 11, no. 9, p. 1162, 2019.
- [21] Y. Wan, J. Tang, and S. Lao, “Distributed conflict-detection and resolution algorithm for UAV swarms based on consensus algorithm and strategy coordination,” *IEEE Access*, vol. 7, pp. 100552–100566, 2019.
- [22] P. Forte, A. Mannucci, H. Andreasson, and F. Pecora, “Online task assignment and coordination in multi-robot fleets,” *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4584–4591, 2021.
- [23] H. Duan, J. Zhao, Y. Deng, Y. Shi, and X. Ding, “Dynamic discrete pigeon-inspired optimization for multi-UAV cooperative search-attack mission planning,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. 57, no. 1, pp. 706–720, 2021.
- [24] K. Hou, Y. Yang, X. Yang, and J. Lai, “Distributed cooperative search algorithm with task assignment and receding horizon predictive control for multiple unmanned aerial vehicles,” *IEEE Access*, vol. 9, pp. 6122–6136, 2021.
- [25] H. Zhang, B. Xin, L.-h. Dou, J. Chen, and K. Hirota, “A review of cooperative path planning of an unmanned aerial vehicle group,” *Frontiers of Information Technology & Electronic Engineering*, vol. 21, no. 12, pp. 1671–1694, 2020.
- [26] R. Luo, H. Zheng, and J. Guo, “Solving the multi-functional heterogeneous UAV cooperative mission planning problem using multi-swarm fruit fly optimization algorithm,” *Sensors (Basel, Switzerland)*, vol. 20, 2020.
- [27] D. A. Egidio, M. Massimiliano, and I. Notaro, “Bi-level flight path planning of UAV formations with collision avoidance,” *Journal of Intelligent & Robotic Systems*, vol. 93, pp. 193–211, 2019.
- [28] A. P. Tirumalai, B. G. Schunck, and R. C. Jain, “Evidential reasoning for building environment maps,” *IEEE Transactions on Systems Man Cybernetics*, vol. 25, no. 1, pp. 10–20, 2002.

Research Article

A Low-Order Partial Integrated Guidance and Control Scheme for Diving Hypersonic Vehicles to Impact Ground Maneuver Target

Tong An ¹, JianHua Wang ², YuLong Pan ¹ and HaiShan Chen ¹

¹People's Liberation Army Air Force Early Warning Academy, Wuhan, Hubei, China

²Department of Aerospace Science and Technology, Space Engineering University, Beijing, China

Correspondence should be addressed to JianHua Wang; wjh_nudt_2013@163.com

Received 2 July 2021; Accepted 26 August 2021; Published 8 September 2021

Academic Editor: Jie Chen

Copyright © 2021 Tong An et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In this article, a low-order partial integrated guidance and control (PIGC) design method is proposed for diving hypersonic vehicles to impact ground maneuver target. A three-channel analytical model of body rates is deduced based on acceleration components of the hypersonic vehicle. By combining the analytical model of body rates and relative dynamic model between the hypersonic vehicle and target, three-channel commands of body rates are directly generated based on the extended state observer (ESO) technique, sliding mode control approach, and dynamic surface control theory in the guidance subsystem. In the attitude control subsystem, a sliding mode controller is designed to track the commands of body rates and generate commands of control surface fin deflections. By making full use of acceleration information of the hypersonic vehicle measured by the mounted accelerometer, the proposed PIGC design method provides a novel solution to compensate the unknown acceleration of the ground maneuver target. Besides, the order of design model is also reduced, and the design process is simplified. The effectiveness and robustness of the PIGC design method are verified and discussed by 6DOF simulation studies.

1. Introduction

The studies on the hypersonic vehicle are increasing due to its unique advantages, such as global rapid attack and ultra-high-speed transportation. Therefore, it has rapidly become a focused research topic in aerospace domain [1]. The dive flight of hypersonic vehicles exhibits characteristics of fast time-varying, strong coupling, and nonlinearity. The coupling between the guidance loop and control loop is strengthened, which proposes higher requirements for the guidance and control (G&C) system design.

Compared with the traditional separated design approach of the G&C system, the integrated guidance and control (IGC) method can fully consider the coupling between centroid motion and rotational motion of the hypersonic vehicle. Besides, the IGC method can ensure the overall stability and improve the accuracy of design model as well as the overall performance of the G&C system [2–4]. Based on the published literatures, the IGC design methods can be divided into three types:

- (1) Single-channel/plane IGC scheme: the three-dimensional (3D) motion of the air vehicle is decomposed into different channels or planes, and the coupling between each channel/plane is regarded as a small amount, which is generally neglected. Hence, the design of the 3D IGC system is reduced to a single-channel or single-plane low-order IGC design process [5–8].
- (2) Full-state IGC scheme: a full-state high-order IGC design model is established, which takes rotational motion model of the air vehicle and relative motion model between the air vehicle and target into account. The order of the design model is usually 8–10, and the design model is generally transformed into a strict feedback form. Then, the backstepping control, dynamic surface control, or other control methods are used to solve the high-order IGC system. Hence, the IGC system design is transformed into a high-order nonlinear control problem [4, 9–14].

- (3) PIGC scheme: the PIGC scheme is executed in the guidance and control loops. The two-loop controller structure is similar to the full-state single-loop controller structure under some conditions. The PIGC method takes body rates instead of acceleration components of the air vehicle as the virtual inputs. The control loop is designed to track the commands of body rates. The order of design model and the number of design parameters can be reduced by using this method [15–17].

The target impacted by air vehicles is often maneuverable, and its acceleration is difficult to acquire directly, which increases the complexity of the G&C system design. In recent years, most studies on the design of the G&C system for air vehicles to impact maneuver target are based on the full-state coupled high-order IGC design model, which contains angles of line-of-sight (LOS), rotational Euler angles of the air vehicle, three-channel body rates, and other motion state variables. The IGC system design is accomplished through the backstepping control method or dynamic surface control method, and the uncertainties containing acceleration components of the target are estimated through some techniques, such as ESO [18–24], sliding mode observer [25, 26], and adaptive law [27]. The effectiveness of this kind of design method has been verified by nonlinear simulation results in the published literatures. However, there are still several disadvantages in the above studies. Firstly, the single-channel/plane IGC system design does not fully consider the coupling between each channel/plane. Secondly, the full-state coupled high-order 3D IGC systems are still based on the traditional separated design method of the G&C system to some extent. The angle of attack, sideslip angle, and bank angle are still the bridges connecting the guidance loop and control loop. The complexity of the high-order design model brings difficulty to the IGC system design. Thirdly, there are still limited studies on the G&C system design for hypersonic vehicles to impact maneuver target.

Taking the above problems into account, a low-order PIGC design method for diving hypersonic vehicles to impact ground maneuver target is proposed in this article. By deducing a three-channel analytical model of body rates based on acceleration components of the hypersonic vehicle, the integration mechanism of rotational motion model of the hypersonic vehicle and relative motion model between the hypersonic vehicle and target is exploited. By combining the analytical model of body rates, relative motion model between the hypersonic vehicle and target, and rotational dynamic model of the hypersonic vehicle, a low-order PIGC scheme is presented, and the design process can be distilled to two steps:

- (1) Three-channel commands of body rates are generated in the guidance loop based on ESO, sliding mode control approach, and dynamic surface control theory.
- (2) Commands of control surface fin deflections are generated in the control loop based on sliding mode control approach.

This novel low-order PIGC scheme can offer an innovative method to compensate the unknown acceleration information of the maneuver target based on the measured acceleration of the hypersonic vehicle and simplify the design process of the G&C system. The content of this article is arranged as follows. Section 2 proposes the 6DOF motion model of the hypersonic vehicle and relative motion model between the hypersonic vehicle and target in dive phase. Section 3.1 deduces the three-channel analytical model of body rates based on acceleration components of the hypersonic vehicle. Section 3.2 presents the design process of the guidance subsystem. Section 3.3 provides the design process of the attitude control subsystem. Section 4 conducts simulation experiments to verify the effectiveness and robustness of the proposed PIGC design method using the generic hypersonic vehicle (GHV) model. Section 5 summarizes the conclusion of this article.

2. Motion Models

The hypersonic vehicle in dive phase has the characteristics of high speed and short flight distance. Hence, the following reasonable assumptions are proposed to simplify the study process: (1) the Earth's rotation is neglected, and the Earth is considered as a flat ground; (2) the mass of the hypersonic vehicle is uniformly distributed, and the influence of body deformation is neglected; (3) the impacts of errors caused by the unmodeled parts of design model are ignored; and (4) the hypersonic vehicle is in unpowered flight during the dive phase regardless of thrust. The denoted motion models in this section consist of centroid and rotational dynamic and kinematical equations of the hypersonic vehicle as well as relative motion equations between the hypersonic vehicle and target. The motion models are given by variables in the ground inertial coordinate system, the body coordinate system, the ballistic coordinate system, and the LOS coordinate system, which are defined in Yan [3].

2.1. Centroid Equations. The centroid dynamic equations of the hypersonic vehicle in the ballistic coordinate system are denoted as follows [3]:

$$\begin{cases} \dot{v} = -g \sin \theta - \frac{D}{m}, \\ \dot{\theta} = \frac{1}{mv} (-mg \cos \theta + L \cos \gamma_V - N \sin \gamma_V), \\ \dot{\sigma} = \frac{1}{mv \cos \theta} (-L \sin \gamma_V - N \cos \gamma_V), \end{cases} \quad (1)$$

where v is the velocity magnitude of the hypersonic vehicle, g is the gravitational acceleration, θ is the flight path angle, σ is the heading angle, γ_V is the bank angle, and m is the mass of the hypersonic vehicle. D , L , and N are the aerodynamic drag, lift, and side force, respectively, which are calculated by

$$\begin{aligned} D &= qSC_D, \\ L &= qSC_L, \\ N &= qSC_N, \end{aligned} \quad (2)$$

where $q = 0.5\rho v^2$ is the dynamic pressure, ρ is the density of atmosphere, C_L is the lift coefficient, C_D is the drag coefficient, C_N is the side force coefficient, and S is the aerodynamic reference area of the hypersonic vehicle.

The centroid kinematical equations of the hypersonic vehicle in the ground inertial coordinate system are given by

$$\begin{cases} \dot{x} = v \cos \theta \cos \sigma, \\ \dot{y} = v \sin \theta, \\ \dot{z} = -v \cos \theta \sin \sigma, \end{cases} \quad (3)$$

where x , y , and z are the components of position vector of the hypersonic vehicle in the ground inertial coordinate system.

2.2. Rotational Equations. The rotational dynamic equations of the hypersonic vehicle in the body coordinate system are denoted as

$$\begin{cases} \dot{\omega}_x = J_x^{-1} M_x + J_x^{-1} (J_y - J_z) \omega_y \omega_z, \\ \dot{\omega}_y = J_y^{-1} M_y + J_y^{-1} (J_z - J_x) \omega_x \omega_z, \\ \dot{\omega}_z = J_z^{-1} M_z + J_z^{-1} (J_x - J_y) \omega_y \omega_x, \end{cases} \quad (4)$$

where ω_x is the roll rate, ω_y is the yaw rate, and ω_z is the pitch rate. J_x , J_y , and J_z are the roll, yaw, and pitch moments of inertia of the hypersonic vehicle, respectively. M_x is the roll moment, M_y is the yaw moment, and M_z is the pitch moment, which are calculated by

$$\begin{aligned} M_x &= qSl_x m_x, \\ M_y &= qSl_y m_y, \\ M_z &= qSl_z m_z, \end{aligned} \quad (5)$$

where l_x , l_y , and l_z are the aerodynamic reference lengths with respect to the roll, yaw, and pitch channels, respectively. m_x is the roll aerodynamic moment coefficient, m_y is the yaw aerodynamic moment coefficient, and m_z is the pitch aerodynamic moment coefficient.

The rotational kinematical equations of the hypersonic vehicle in the body coordinate system are provided by

$$\begin{cases} \dot{\varphi} = \omega_y \sin \gamma + \omega_z \cos \gamma, \\ \dot{\psi} = (\omega_y \cos \gamma - \omega_z \sin \gamma) \sec \varphi, \\ \dot{\gamma} = \omega_x - (\omega_y \cos \gamma - \omega_z \sin \gamma) \tan \varphi, \end{cases} \quad (6)$$

where φ is the pitch angle, ψ is the yaw angle, and γ is the roll angle.

2.3. Relative Motion Equations. Figure 1 illustrates the schematic diagram of relative motion between the hypersonic vehicle and ground maneuver target. O_B and T represent the centroids of the hypersonic vehicle and target, respectively. $O - XYZ$ is the ground inertial coordinate

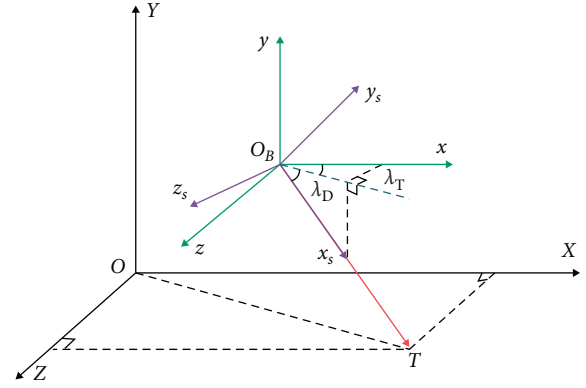


FIGURE 1: Schematic diagram of relative motion.

system. $O_B - x_s y_s z_s$ is the LOS coordinate system. The relative motion equations between the hypersonic vehicle and target are denoted as [3]

$$\begin{cases} \ddot{r} = r\dot{\lambda}_D^2 + r\dot{\lambda}_T^2 \cos^2 \lambda_D + a_{xs}^T - a_{xs}^V, \\ \ddot{\lambda}_D = \frac{(-2r\dot{\lambda}_D - r\dot{\lambda}_T^2 \cos \lambda_D \sin \lambda_D)}{r} + \frac{(a_{ys}^T - a_{ys}^V)}{r}, \\ \ddot{\lambda}_T = \frac{(2r\dot{\lambda}_D \dot{\lambda}_T \sin \lambda_D - 2r\dot{\lambda}_T \cos \lambda_D)}{(r \cos \lambda_D)} - \frac{(a_{zs}^T - a_{zs}^V)}{(r \cos \lambda_D)}, \end{cases} \quad (7)$$

where r is the relative distance between the hypersonic vehicle and target, λ_D is the elevation angle of LOS, λ_T is the azimuth angle of LOS, a_{xs}^V , a_{ys}^V , and a_{zs}^V are the acceleration components of the hypersonic vehicle in the LOS coordinate system, and a_{xs}^T , a_{ys}^T , and a_{zs}^T are the acceleration components of the target in the LOS coordinate system.

3. Partial Integrated Guidance and Control System Design

In this section, the design scheme of a novel PIGC system is presented, which contains a guidance loop and a control loop. Firstly, a three-channel analytical model of body rates is deduced. Then, three-channel commands of body rates are generated in the guidance loop based on ESO, sliding mode control approach, and dynamic surface control theory. Finally, a sliding mode attitude controller is designed to track the commands of body rates, and commands of control surface fin deflections are directly obtained in the control loop.

3.1. Three-Channel Analytical Model of Body Rates.

According to equation (3), the velocity vectors of the hypersonic vehicle denoted in the ground inertial coordinate system and the body coordinate system, respectively, hold the following relation:

$$\begin{bmatrix} v \cos \theta \cos \sigma \\ v \sin \theta \\ -v \cos \theta \sin \sigma \end{bmatrix} = \mathbf{T}_{GB} \begin{bmatrix} u \\ v \\ w \end{bmatrix}, \quad (8)$$

where $[u \ v \ w]^T$ is the velocity vector of the hypersonic vehicle in the body coordinate system and \mathbf{T}_{GB} is the transition matrix from the body coordinate system to the ground inertial coordinate system. The derivatives of both sides of equation (8) can be written as

$$\frac{d}{dt} \begin{pmatrix} v \cos \theta \cos \sigma \\ v \sin \theta \\ -v \cos \theta \sin \sigma \end{pmatrix} = \frac{d}{dt} (\mathbf{T}_{GB}) \begin{bmatrix} u \\ v \\ w \end{bmatrix} + \mathbf{T}_{GB} \frac{d}{dt} \begin{pmatrix} u \\ v \\ w \end{pmatrix}. \quad (9)$$

The left side of equation (9) holds the following relation:

$$\frac{d}{dt} \begin{pmatrix} v \cos \theta \cos \sigma \\ v \sin \theta \\ -v \cos \theta \sin \sigma \end{pmatrix} = \mathbf{T}_{GH} \begin{bmatrix} a_{xh}^V \\ a_{yh}^V \\ a_{zh}^V \end{bmatrix} = \mathbf{T}_{GH} \begin{bmatrix} \dot{v} \\ v\dot{\theta} \\ -v \cos \theta \dot{\sigma} \end{bmatrix}, \quad (10)$$

where \mathbf{T}_{GH} is the transition matrix from the ballistic coordinate system to the ground inertial coordinate system and a_{xh}^V , a_{yh}^V , and a_{zh}^V are the acceleration components of the hypersonic vehicle along the three axes in the ballistic coordinate system. The first term on the right side of equation (9) can be calculated by

$$\frac{d}{dt} (\mathbf{T}_{GB}) \begin{bmatrix} u \\ v \\ w \end{bmatrix} = \begin{bmatrix} T_{11} & T_{12} & T_{13} \\ T_{21} & T_{22} & T_{23} \\ T_{31} & T_{32} & T_{33} \end{bmatrix} \begin{bmatrix} \dot{\varphi} \\ \dot{\psi} \\ \dot{\gamma} \end{bmatrix}, \quad (11)$$

where

$$\begin{aligned} T_{11} &= -\sin \varphi \cos \psi u - \cos \varphi \cos \psi \cos \gamma v + \cos \varphi \cos \psi \sin \gamma w, \\ T_{12} &= -\cos \varphi \sin \psi u + \cos \psi \sin \gamma v + \sin \varphi \sin \psi \cos \gamma w \\ &\quad + \cos \psi \cos \gamma w - \sin \varphi \sin \psi \sin \gamma w, \\ T_{13} &= \sin \psi \cos \gamma v + \sin \varphi \cos \psi \sin \gamma v - \sin \psi \sin \gamma w + \sin \varphi \cos \psi \cos \gamma w, \\ T_{21} &= \cos \varphi u - \sin \varphi \cos \gamma v + \sin \varphi \sin \gamma w, \\ T_{22} &= 0, \\ T_{23} &= -\cos \varphi \sin \gamma v - \cos \varphi \cos \gamma w, \\ T_{31} &= \sin \varphi \sin \psi u + \cos \varphi \sin \psi \cos \gamma v - \cos \varphi \sin \psi \sin \gamma w, \\ T_{32} &= -\cos \varphi \cos \psi u - \sin \psi \sin \gamma v + \sin \varphi \cos \psi \cos \gamma v \\ &\quad - \sin \psi \cos \gamma w - \sin \varphi \cos \psi \sin \gamma w, \\ T_{33} &= \cos \psi \cos \gamma v - \sin \varphi \sin \psi \sin \gamma v - \cos \psi \sin \gamma w - \sin \varphi \sin \psi \cos \gamma w. \end{aligned} \quad (12)$$

Combining equations (6) and (9)–(11), the following equation can be obtained:

$$\mathbf{T}_{GH} \begin{bmatrix} a_{xh}^V \\ a_{yh}^V \\ a_{zh}^V \end{bmatrix} = \mathbf{T}_{GB} \begin{bmatrix} \dot{u} \\ \dot{v} \\ \dot{w} \end{bmatrix} + \mathbf{AB} \begin{bmatrix} \omega_x \\ \omega_y \\ \omega_z \end{bmatrix}, \quad (13)$$

where $\mathbf{A}, \mathbf{B} \in \mathbf{R}^{3 \times 3}$ are given by

$$\mathbf{A} = \begin{bmatrix} T_{11} & T_{12} & T_{13} \\ T_{21} & T_{22} & T_{23} \\ T_{31} & T_{32} & T_{33} \end{bmatrix}, \quad (14)$$

$$\mathbf{B} = \begin{bmatrix} 0 & \sin \gamma & \cos \gamma \\ 0 & \cos \gamma \sec \varphi & -\sin \gamma \sec \varphi \\ 1 & -\cos \gamma \tan \varphi & \sin \gamma \tan \varphi \end{bmatrix}.$$

According to equation (13), a three-channel analytical model of body rates can be obtained:

$$\begin{bmatrix} \omega_x \\ \omega_y \\ \omega_z \end{bmatrix} = (\mathbf{AB})^{-1} \left(\mathbf{T}_{GH} \begin{bmatrix} a_{xh}^V \\ a_{yh}^V \\ a_{zh}^V \end{bmatrix} - \mathbf{T}_{GB} \begin{bmatrix} \dot{u} \\ \dot{v} \\ \dot{w} \end{bmatrix} \right). \quad (15)$$

Based on the expected acceleration components of the hypersonic vehicle along the three axes in the ballistic coordinate system, the corresponding desired body rates of the hypersonic vehicle with respect to the roll, yaw, and pitch channels can be generated by equation (15). This analytical equation can help in replacing the process of “generating commands of angle of attack and sideslip angle/bank angle according to the desired acceleration of the hypersonic vehicle \rightarrow designing the attitude control system to track the commands of rotational Euler angles” in the traditional design scheme of the G&C system. Besides, the integration of centroid motion model and rotational motion model can be improved, and the design procedure of the G&C system can be simplified.

3.2. Guidance Subsystem Design. Equation (7) can be rewritten as

$$\begin{bmatrix} \ddot{\lambda}_D \\ \ddot{\lambda}_T \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} + \begin{bmatrix} \frac{-1}{r} & 0 \\ 0 & \frac{1}{(r \cos \lambda_D)} \end{bmatrix} \begin{bmatrix} a_{ys}^V \\ a_{zs}^V \end{bmatrix}, \quad (16)$$

where

$$\begin{bmatrix} f_1 \\ f_2 \end{bmatrix} = \begin{bmatrix} \frac{(-2r\dot{\lambda}_D - r\dot{\lambda}_T^2 \cos \lambda_D \sin \lambda_D + a_{ys}^T)}{r} \\ \frac{(2r\dot{\lambda}_D \dot{\lambda}_T \sin \lambda_D - 2r\dot{\lambda}_T \cos \lambda_D - a_{zs}^T)}{(r \cos \lambda_D)} \end{bmatrix}. \quad (17)$$

In this article, the hypersonic vehicle is steered to the ground maneuver target by zeroing $\dot{\lambda}_D$ and $\dot{\lambda}_T$. Based on equation (17), a_{ys}^V and a_{zs}^V are regarded as the virtual inputs to zero the angle rates of LOS. A sliding mode surface vector is conducted as $\mathbf{S}_1 = [S_D \ S_T]^T = [\dot{\lambda}_D \ \dot{\lambda}_T]^T$, and the anticipant dynamic of \mathbf{S}_1 is chosen as

$$\begin{bmatrix} \dot{S}_D \\ \dot{S}_T \end{bmatrix} = \begin{bmatrix} -\varepsilon_D |S_D|^{\rho_D} \text{sgn}(S_D) - k_D S_D \\ -\varepsilon_T |S_T|^{\rho_T} \text{sgn}(S_T) - k_T S_T \end{bmatrix}, \quad (18)$$

where ε_D , ε_T , ρ_D , ρ_T , k_D , and k_T are parameters that need to be designed suitably to make S_D and S_T converge to zero gradually without chattering phenomenon.

In equation (16), f_1 and f_2 contain acceleration components of the ground maneuver target, which cannot be directly obtained by the seeker of the hypersonic vehicle. In this article, the acceleration components of the target in the LOS coordinate system are recognized as bounded uncertainties. The ESO technology is used to estimate f_1 and f_2 including uncertainties [28]:

$$\begin{cases} \dot{\mathbf{Z}}_1 = \mathbf{Z}_2 - \beta_1 \left(\mathbf{Z}_1 - \begin{bmatrix} \dot{\lambda}_D \\ \dot{\lambda}_T \end{bmatrix} \right) + \begin{bmatrix} \frac{-1}{r} & 0 \\ 0 & \frac{1}{(r \cos \lambda_D)} \end{bmatrix} \begin{bmatrix} a_{ys}^V \\ a_{zs}^V \end{bmatrix}, \\ \dot{\mathbf{Z}}_2 = \beta_2 \left(\mathbf{Z}_1 - \begin{bmatrix} \dot{\lambda}_D \\ \dot{\lambda}_T \end{bmatrix} \right), \end{cases} \quad (19)$$

where β_1 and β_2 are parameters that need to be designed, \mathbf{Z}_1 , $\mathbf{Z}_2 \in \mathbf{R}^{2 \times 1}$ are the state vectors of the ESO, and the two elements in \mathbf{Z}_2 are the estimated values of f_1 and f_2 . In practical engineering application, real-time a_{ys}^V and a_{zs}^V can be obtained by

$$\begin{bmatrix} a_{ys}^V \\ a_{zs}^V \end{bmatrix} = \mathbf{T}_{SG,23} \left(\mathbf{T}_{GB} \begin{bmatrix} a_{xb}^V \\ a_{yb}^V \\ a_{zb}^V \end{bmatrix} + \mathbf{AB} \begin{bmatrix} \omega_x \\ \omega_y \\ \omega_z \end{bmatrix} \right), \quad (20)$$

where $\mathbf{T}_{SG,23}$ represents the matrix reconstructed with the second and third rows of the transition matrix from the ground inertial coordinate system to the LOS coordinate system:

$$\mathbf{T}_{SG,23} = \begin{bmatrix} -\sin \lambda_D \cos \lambda_T & \cos \lambda_D \sin \lambda_D \sin \lambda_T \\ \sin \lambda_T & 0 & \cos \lambda_T \end{bmatrix}, \quad (21)$$

where $[a_{xb}^V \ a_{yb}^V \ a_{zb}^V]^T$ is the acceleration of the hypersonic vehicle in the body coordinate system and is measured by the accelerometer mounted on the hypersonic vehicle. Hence, f_1 and f_2 can be estimated and the acceleration components of the target in the LOS coordinate system can be well compensated by making full use of acceleration information of the hypersonic vehicle. It should be noted that real-time a_{ys}^V and a_{zs}^V in the simulation experiments are obtained based on the centroid dynamic model of the hypersonic vehicle in this article.

Combining equations (16), (18), and (19), the desired virtual inputs can be obtained by

$$\begin{bmatrix} a_{ys}^{V,c} \\ a_{zs}^{V,c} \end{bmatrix} = \begin{bmatrix} \frac{-1}{r} & 0 \\ 0 & \frac{1}{(r \cos \lambda_D)} \end{bmatrix}^{-1} \left(\begin{bmatrix} \dot{S}_D \\ \dot{S}_T \end{bmatrix} - \mathbf{Z}_2 \right). \quad (22)$$

Let $\mathbf{C} = \mathbf{T}_{SG,23} \mathbf{T}_{GH}$. Combining equations (15) and (22), the desired three-channel body rates of the hypersonic vehicle can be directly generated as

$$\begin{bmatrix} \omega_{x,c} \\ \omega_{y,c} \\ \omega_{z,c} \end{bmatrix} = (\mathbf{AB})^{-1} \left(\mathbf{T}_{GH} \mathbf{C}^{-1} \begin{bmatrix} \frac{-1}{r} & 0 \\ 0 & \frac{1}{(r \cos \lambda_D)} \end{bmatrix}^{-1} \left(\begin{bmatrix} \dot{S}_D \\ \dot{S}_T \end{bmatrix} - \mathbf{Z}_2 \right) - \mathbf{T}_{GB} \begin{bmatrix} \dot{u} \\ \dot{v} \\ \dot{w} \end{bmatrix} \right). \quad (23)$$

It should be noted that the change rates of the desired three-channel body rates are indispensable in the design process of the attitude control subsystem. However, differentiating the desired body rates directly is more inconvenient in engineering application and increases the amount of calculation. Hence, based on the dynamic surface control theory, $[\omega_{x,c} \ \omega_{y,c} \ \omega_{z,c}]^T$ is passed through a low-pass filter:

$$\tau \begin{bmatrix} \dot{\omega}_{x,d} \\ \dot{\omega}_{y,d} \\ \dot{\omega}_{z,d} \end{bmatrix} + \begin{bmatrix} \omega_{x,d} \\ \omega_{y,d} \\ \omega_{z,d} \end{bmatrix} = \begin{bmatrix} \omega_{x,c} \\ \omega_{y,c} \\ \omega_{z,c} \end{bmatrix}, \quad (24)$$

where $\tau \in \mathbf{R}^{3 \times 3}$ is the designed parameter matrix of the low-pass filter. The outputs $\omega_{x,d}$, $\omega_{y,d}$, and $\omega_{z,d}$ are the final three-channel commands of body rates generated in the guidance subsystem. By designing the attitude controller to track the three-channel commands of body rates, λ_D and λ_T can gradually converge to zero. Hence, the hypersonic vehicle can successfully impact the ground maneuver target.

It should be noted that the inverse of matrix \mathbf{AB} and the generalized inverse of matrix \mathbf{C} are included in equation (23). The elements of matrix \mathbf{AB} are the functions of components of velocity vector of the hypersonic vehicle in the body coordinate system and rotational Euler angles. The elements of matrix \mathbf{C} are the functions of angles of LOS, flight path angle, and heading angle. As the two matrixes cannot be proved to be invertible/generalized invertible theoretically, the inverse of \mathbf{AB} and the generalized inverse of \mathbf{C} are assumed to exist in the whole dive phase. This assumption can only be validated by 6DOF simulation results.

3.3. Attitude Control Subsystem Design. To track the commands of body rates generated in the guidance subsystem, the commanded control surface fin deflections of the hypersonic vehicle should be calculated by designing the attitude control subsystem. To simplify the design process, equation (4) is rewritten as

$$\begin{bmatrix} \dot{\omega}_x \\ \dot{\omega}_y \\ \dot{\omega}_z \end{bmatrix} = \mathbf{D} + \mathbf{E} \begin{bmatrix} \delta_a \\ \delta_e \\ \delta_r \end{bmatrix}, \quad (25)$$

where

$$\mathbf{D} = \begin{bmatrix} \dot{\omega}_x^* \\ \dot{\omega}_y^* \\ \dot{\omega}_z^* \end{bmatrix} = \begin{bmatrix} J_x^{-1}(J_y - J_z)\omega_z\omega_y + J_x^{-1}qSl_x m_x^* \\ J_y^{-1}(J_z - J_x)\omega_x\omega_z + J_y^{-1}qSl_y m_y^* \\ J_z^{-1}(J_x - J_y)\omega_y\omega_x + J_z^{-1}qSl_z m_z^* \end{bmatrix}, \quad (26)$$

$$\mathbf{E} = \begin{bmatrix} J_x^{-1}qSl_x & & \\ & J_y^{-1}qSl_y & \\ & & J_z^{-1}qSl_z \end{bmatrix} \begin{bmatrix} m_{x,1} & m_{x,2} & m_{x,3} \\ m_{y,1} & m_{y,2} & m_{y,3} \\ m_{z,1} & m_{z,2} & m_{z,3} \end{bmatrix}. \quad (27)$$

In equation (25), δ_a , δ_e , and δ_r are the right elevon, left elevon, and rudder fin deflections of the GHV, respectively. In equation (27), $m_{i,j}$, $i = x, y, z$, $j = 1, 2, 3$ are the coefficients related to first-degree control surface fin deflections.

In equation (26), m_x^* , m_y^* , and m_z^* represent the remaining terms of m_x , m_y , and m_z after removing the relevant parts of first-degree control surface fin deflections.

The sliding mode control method is used in the attitude control subsystem. The sliding mode surface vector is selected as

$$\mathbf{S}_2 = \begin{bmatrix} S_x \\ S_y \\ S_z \end{bmatrix} = \begin{bmatrix} \omega_x - \omega_{x,d} \\ \omega_y - \omega_{y,d} \\ \omega_z - \omega_{z,d} \end{bmatrix}. \quad (28)$$

The desired dynamic of \mathbf{S}_2 is selected as

$$\begin{bmatrix} \dot{S}_x \\ \dot{S}_y \\ \dot{S}_z \end{bmatrix} = \begin{bmatrix} -\varepsilon_x |S_x|^{\rho_x} \text{sgn}(S_x) - k_x S_x \\ -\varepsilon_y |S_y|^{\rho_y} \text{sgn}(S_y) - k_y S_y \\ -\varepsilon_z |S_z|^{\rho_z} \text{sgn}(S_z) - k_z S_z \end{bmatrix}, \quad (29)$$

where $\varepsilon_i, \rho_i, k_i$, $i = x, y, z$ are parameters that need to be designed. For the attitude control subsystem, the significance of finite-time convergence is obvious. Hence, the values of $\varepsilon_i, \rho_i, k_i$, $i = x, y, z$ should be set reasonably to ensure the high convergence rate of \mathbf{S}_2 and avoid the occurrence of chattering.

Combing equations (25), (28), and (29), three-channel commands of control surface fin deflections are obtained by

$$\begin{bmatrix} \delta_{a,c} \\ \delta_{e,c} \\ \delta_{r,c} \end{bmatrix} = \mathbf{E}^{-1} \left(\begin{bmatrix} \dot{S}_x \\ \dot{S}_y \\ \dot{S}_z \end{bmatrix} + \begin{bmatrix} \dot{\omega}_{x,d} \\ \dot{\omega}_{y,d} \\ \dot{\omega}_{z,d} \end{bmatrix} - \mathbf{D} \right). \quad (30)$$

For the convenience of calculating the current commands of control surface fin deflections, $[\dot{\omega}_x^* \ \dot{\omega}_y^* \ \dot{\omega}_z^*]^T$ is obtained based on the commands of control surface fin deflections generated one time step before. Define $\mathbf{x}_1 = [\lambda_D \ \lambda_T]^T$, $\mathbf{x}_1 \in \mathbf{R}^{2 \times 1}$, $\mathbf{x}_2 = [\omega_x \ \omega_y \ \omega_z]^T$, $\mathbf{x}_2 \in \mathbf{R}^{3 \times 1}$, $\mathbf{u} = [\delta_a \ \delta_e \ \delta_r]^T$, $\mathbf{u} \in \mathbf{R}^{3 \times 1}$. Based on the above design process of guidance and attitude control subsystems, the design model of the PIGC system is given by

$$\begin{cases} \dot{\mathbf{x}}_1 = \mathbf{F} + \mathbf{G} \cdot \mathbf{x}_2, \\ \dot{\mathbf{x}}_2 = \mathbf{D} + \mathbf{E} \cdot \mathbf{u}, \end{cases} \quad (31)$$

where

$$\mathbf{F} = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} + \begin{bmatrix} -\frac{1}{r} & 0 \\ 0 & \frac{1}{(r \cos \lambda_D)} \end{bmatrix} \mathbf{T}_{SG,23} \mathbf{T}_{GB} \begin{bmatrix} \dot{u} \\ \dot{v} \\ \dot{w} \end{bmatrix}, \quad (32)$$

$$\mathbf{G} = \begin{bmatrix} -\frac{1}{r} & 0 \\ 0 & \frac{1}{(r \cos \lambda_D)} \end{bmatrix} \mathbf{T}_{SG,23} \mathbf{AB}.$$

It is noted that the order of the design model is only 5. Besides, the process of calculating the commands of rotational Euler angles based on the desired overload of the hypersonic vehicle in the traditional G&C system design is elided. Hence, the design process of the proposed PIGC system exhibits the characteristic of simplification.

Taking the response delay of fin actuators into account, the generated commands of control surface fin deflections are regarded as the inputs of the second-order fin actuators. The outputs are regarded as the actual fin deflections applied to the hypersonic vehicle. The model of the fin actuators is given by

$$\frac{d\dot{\delta}_i}{dt} = -\omega_n^2 \delta_i - 2\xi\omega_n \dot{\delta}_i + \omega_n^2 \delta_{i,c}, \quad i = a, e, r, \quad (33)$$

where $\omega_n = 20$ Hz is the natural frequency of the second-order model and $\xi = 0.7$ is the damping ratio of the model.

4. Simulation Tests and Discussion

4.1. Parameter Setting. The effectiveness and robustness of the proposed PIGC method are validated by 6DOF simulation experiments based on the GHV model. The detailed model parameters and aerodynamic force and moment coefficients of GHV are provided in [29]. The maximum change rate of fin deflections is 100 deg/s, and the magnitudes of fin deflections are constrained as $-20^\circ \leq \delta_i \leq +20^\circ$, $i = a, e, r$. The initial values of 6DOF motion states of the diving hypersonic vehicle are presented in Table 1. The values of design parameters are given in Table 2.

The ground maneuver target is considered as a point mass, and the centroid motion model is given by

$$\begin{cases} [\dot{x}_T & \dot{y}_T & \dot{z}_T]^T = [v_{xT} & v_{yT} & v_{zT}]^T, \\ [\dot{v}_{xT} & \dot{v}_{yT} & \dot{v}_{zT}]^T = [a_{xT} & a_{yT} & a_{zT}]^T, \end{cases} \quad (34)$$

where $[x_T \ y_T \ z_T]^T$, $[v_{xT} \ v_{yT} \ v_{zT}]^T$, and $[a_{xT} \ a_{yT} \ a_{zT}]^T$ are the position, velocity, and acceleration vectors of the target in the ground inertial coordinate system. The initial components of position vector and velocity vector of the target are set as $x_{T0} = 120$ km, $y_{T0} = 0$ km, $z_{T0} = 30$ km, $v_{xT0} = v_{yT0} = 0$ m/s, and $v_{zT0} = -25$ m/s. The acceleration components of the target are set as $a_{xT} = 2 \cos(0.01t)$ m/s², $a_{yT} = 0$ m/s², and $a_{zT} = 1$ m/s².

In this article, the longitude, latitude, and altitude of the origin of the ground inertial coordinate system are set as 0 deg, 0 deg, and 0 m. The ox axis is set to the straight east, and the oy axis is set to be vertical to the ground and upward. When the altitude of the hypersonic vehicle is less than zero, the simulation test is terminated. At the terminal time, the distance between the hypersonic vehicle and target is the miss distance.

4.2. Effectiveness Verification. In order to verify the effectiveness and robustness of the proposed PIGC design method, the nominal values of aerodynamic force coefficients, aerodynamic moment coefficients, and density of atmosphere are multiplied by ① $1 + 0.2 \sin(t)$, ② $1 + 0.3 \sin(t)$, and ③

TABLE 1: Initial values of 6DOF motion states of the hypersonic vehicle.

States	Values	States	Values
v_0	2600 m/s	φ_0	10°
θ_0	0°	ψ_0	-5°
σ_0	0°	γ_0	-40°
x_0	0 m	ω_{x0}	$0^\circ/\text{s}$
y_0	27000 m	ω_{y0}	$0^\circ/\text{s}$
z_0	0 m	ω_{z0}	$0^\circ/\text{s}$

TABLE 2: Values of design parameters of the PIGC system.

Parameters	Values	Parameters	Values
ε_D	0.03	$\varepsilon_x, \varepsilon_y, \varepsilon_z$	0.002
ε_T	0.015	k_x, k_y, k_z	1
k_D	1.5	ρ_x, ρ_y, ρ_z	0.6
k_T	0.75	β_1	200
ρ_D, ρ_T	0.6	β_2	10000
τ	diag(0.04, 0.04, 0.04)		

$1 + 0.4 \sin(t)$ (t stands for the simulation time). Simulation studies are conducted with the actual aerodynamic force and moment coefficients and atmospheric density deviating as ①–③. Simulation results of the three conditions are represented by red solid lines, green dash lines, and blue dot lines, respectively, as shown in Figures 2–8.

Figure 2 presents the curves of velocity, flight path angle, and heading angle of the hypersonic vehicle in the dive phase under different deviations. As shown in Figure 2(a), the velocity magnitude of the hypersonic vehicle gradually and smoothly decreases under each deviation. It is also observed that the flight path angle and heading angle vary smoothly in the total dive phase, as illustrated in Figures 2(b) and 2(c).

Figure 3 illustrates the 3D flight trajectories of the hypersonic vehicle and curves of relative distances between the hypersonic vehicle and target under different deviations. A red pentagram indicates the location of the ground maneuver target when the simulation terminates (considering that the distances between the terminal locations of the ground maneuver target under different deviations are very short, the terminal location of the target under deviation ① is taken as an example). It is indicated that the flight trajectories almost coincided with each other under different deviations. For deviations ①–③, the miss distances at the terminal time are 16.07 m, 22.55 m, and 31.06 m, respectively. Considering the reference length and high flight speed of the hypersonic vehicle, the miss distances under different deviations are acceptable and the proposed PIGC design method can accurately steer the hypersonic vehicle to impact the ground maneuver target with good robustness.

Figure 4 illustrates the curves of elevation and azimuth angles of LOS and change rates of angles of LOS (sliding mode surfaces S_D and S_T) under different deviations. It can be observed that when the simulation time is more than 30 seconds, the change rates of elevation and azimuth angles of LOS almost converge to zero. It should be noted that when the simulation is near the end, the elevation and azimuth angles of LOS drastically change. This phenomenon is due to

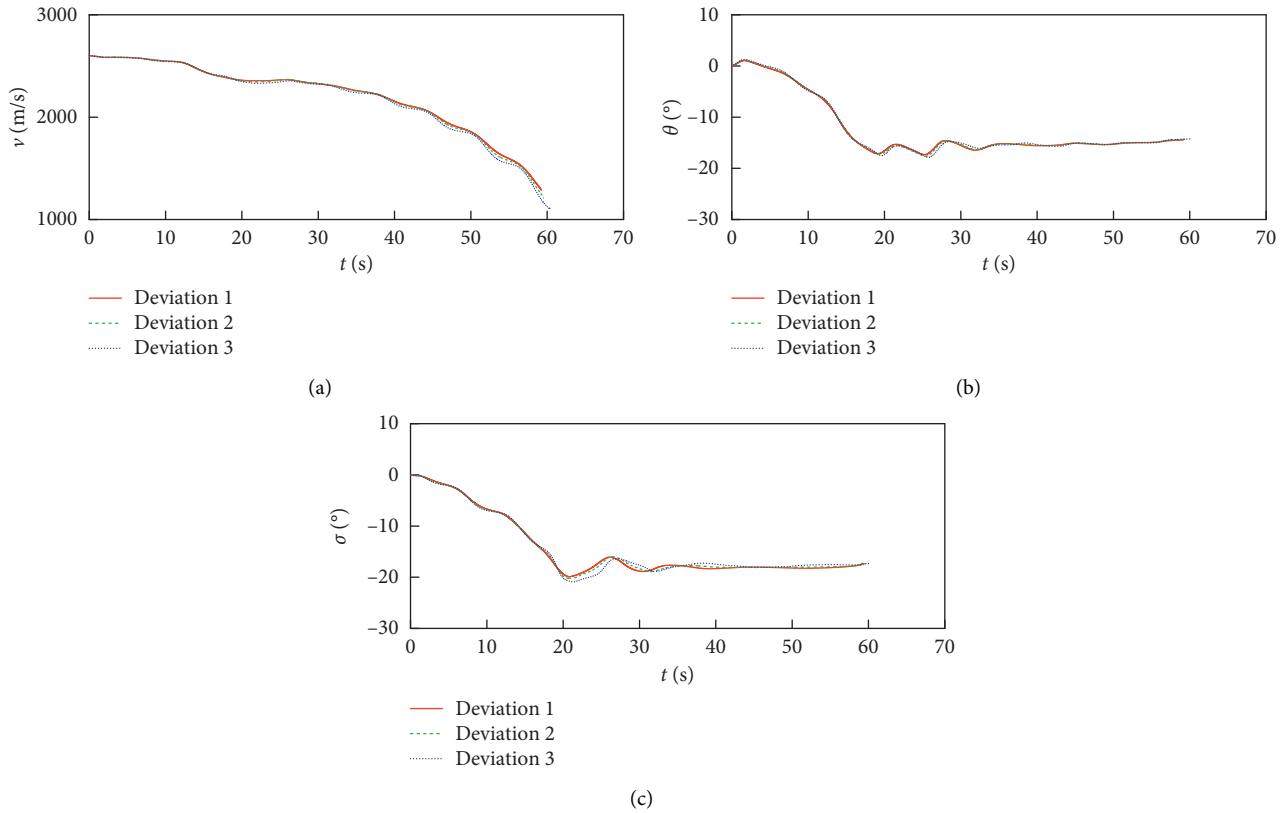


FIGURE 2: Curves of velocity, flight path angle, and heading angle.

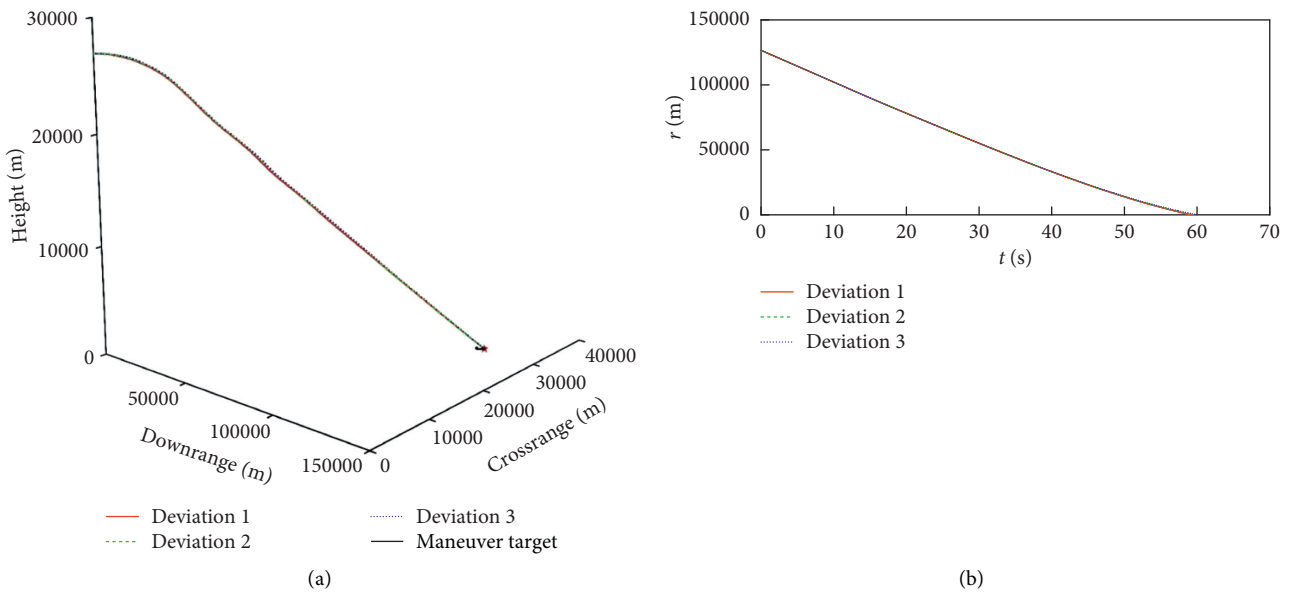


FIGURE 3: Curves of flight trajectories of the hypersonic vehicle and relative distances between the hypersonic vehicle and target.

the fact that the two angles are calculated based on the relative spatial position of the hypersonic vehicle and target. When the distance between the hypersonic vehicle and target reduces to a small amount, the two angles drastically change.

Figures 5–7 illustrate the curves of rotational motion state variables of the hypersonic vehicle under different deviations. It is found that all attitude angles and body rates change boundedly and smoothly in the dive phase, which indicates that the attitude of the hypersonic vehicle is

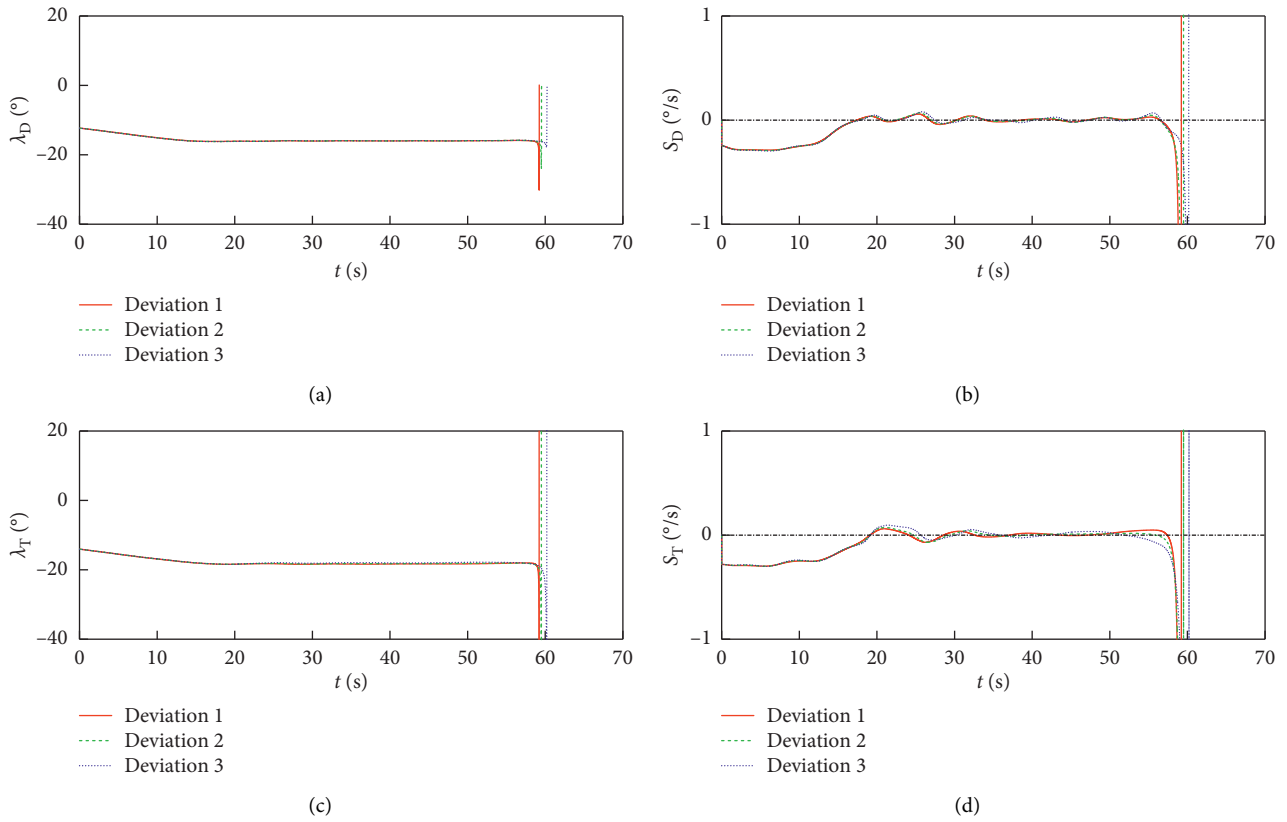


FIGURE 4: Curves of elevation and azimuth angles of LOS and change rates of angles of LOS.

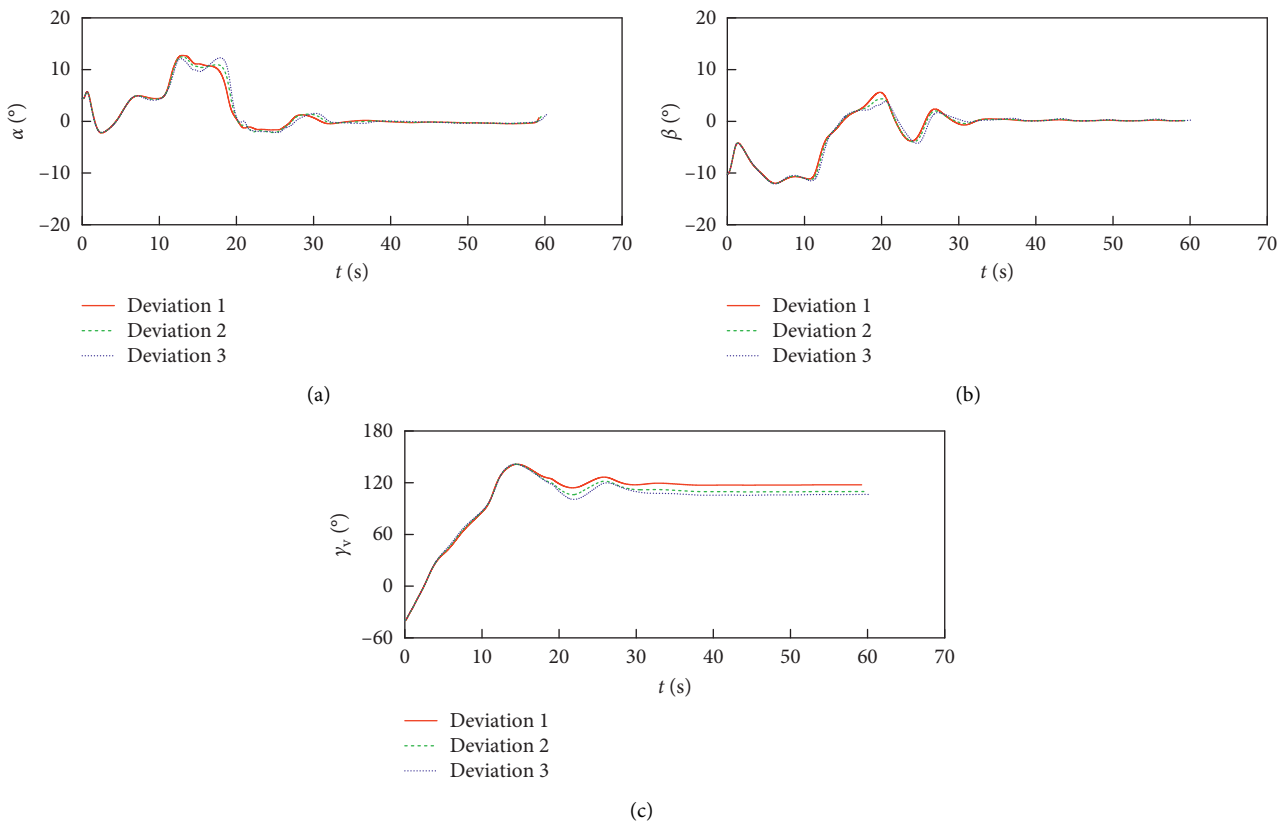


FIGURE 5: Curves of angle of attack, sideslip angle, and bank angle.

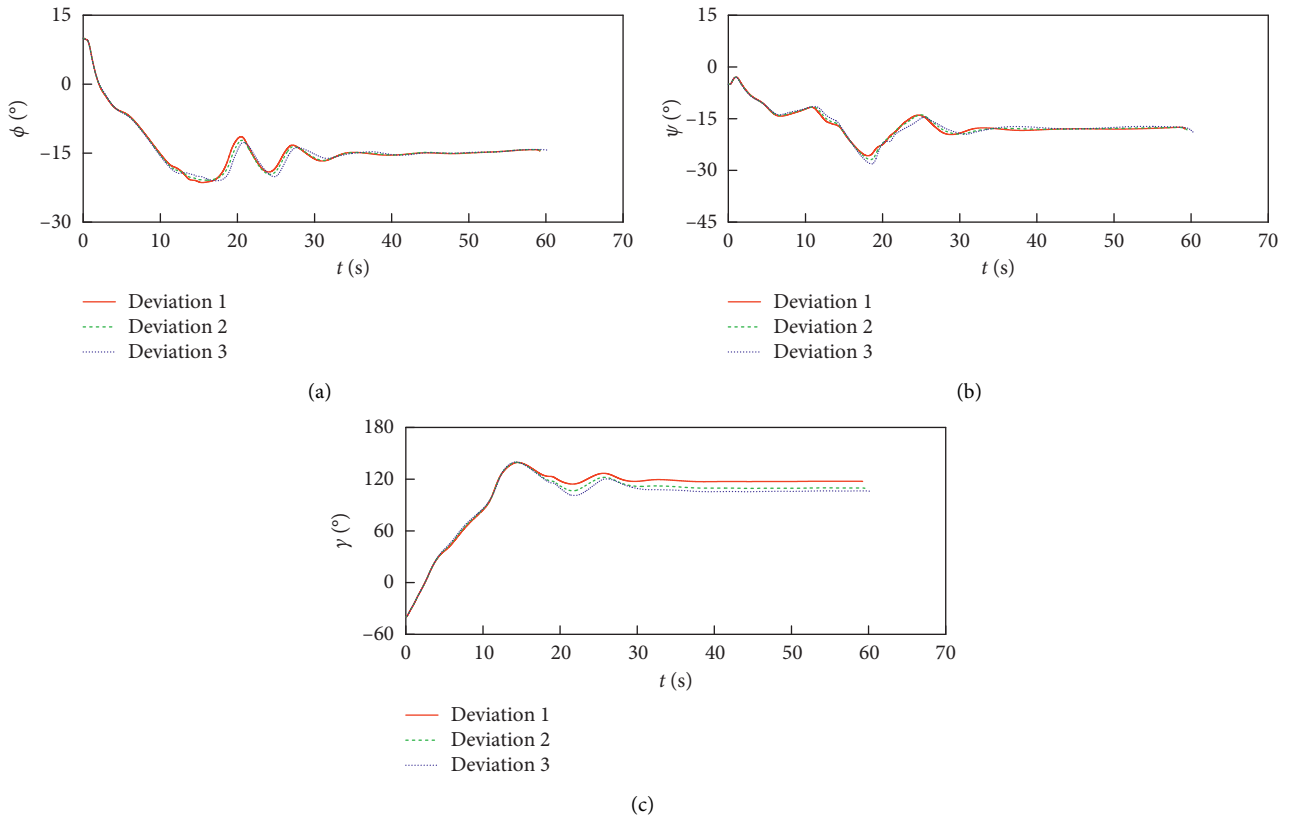


FIGURE 6: Curves of pitch angle, yaw angle, and roll angle.

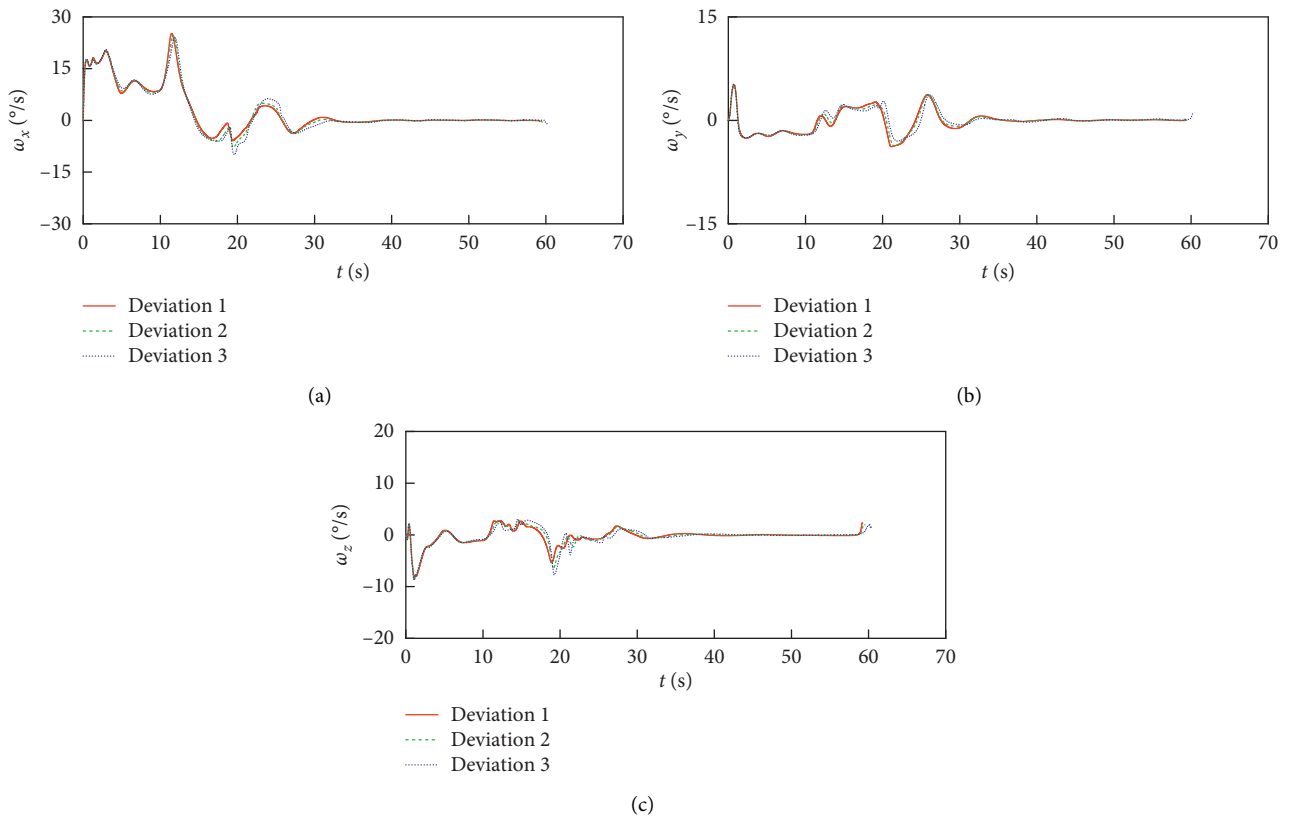


FIGURE 7: Curves of roll rate, yaw rate, and pitch rate.

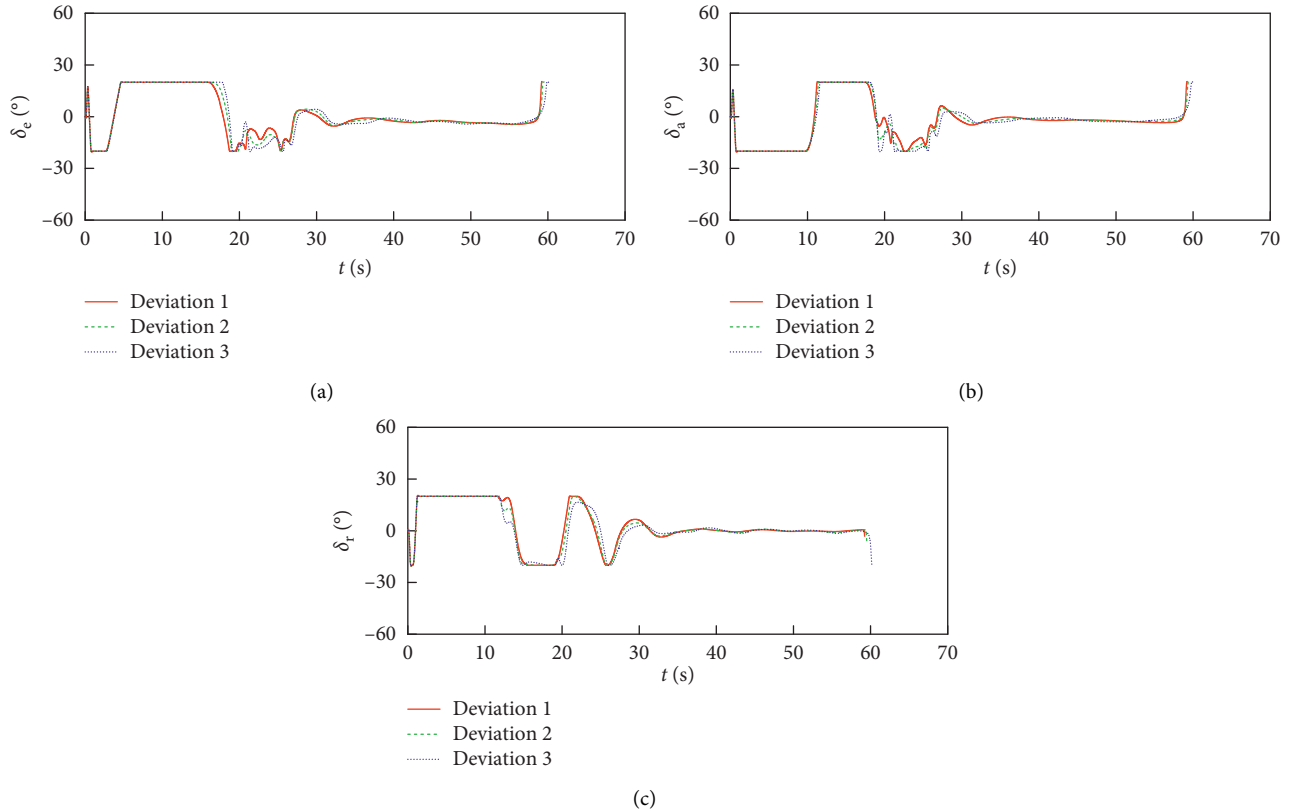


FIGURE 8: Curves of actual fin deflections.

controllable under each deviation. However, the change rates of rotational motion state variables are larger in the first 30 seconds of the simulation. This phenomenon is caused by the large deviation between the initial flight direction of the hypersonic vehicle and initial location of the ground maneuver target. The phenomenon is also in accordance with the fact that the elevation and azimuth angles of LOS slowly decrease in the first 30 seconds as shown in Figure 4. The three-channel body rates also exhibit similar characteristics as mentioned above.

Figure 8 illustrates the curves of actual control surface fin deflections applied to the hypersonic vehicle under different deviations, which are smooth and bounded. It is indicated that the hypersonic vehicle can complete the G&C mission within the ability of actuators under different deviations. In addition, the assumption that the inverse of matrix \mathbf{AB} and the generalized inverse of matrix \mathbf{C} in equation (23) exist during the whole dive phase is validated by the simulation results.

5. Conclusion

In this article, a low-order PIGC method for diving hypersonic vehicles to impact ground maneuver target with good robustness is designed. A three-channel analytical model of body rates based on acceleration components of the hypersonic vehicle is deduced. By combining the analytical model of body rates and relative motion model between the hypersonic vehicle and target, three-channel commands of body rates are directly generated based on

ESO, sliding mode control approach, and dynamic surface control theory. A sliding mode controller is designed in the attitude control subsystem to track the commands of body rates. The acceleration components of the maneuver target in the relative motion model can be compensated by ESO with the use of acceleration information of the hypersonic vehicle. Hence, the utilization efficiency of the acceleration sensor mounted on the hypersonic vehicle is improved, and the impact accuracy can be guaranteed. Besides, the order of design model is reduced, and the design process of the G&C system is simplified. However, this article is aimed at proposing a novel PIGC scheme and verifying its effectiveness and robustness. Hence, the influences of uncertainties brought by dynamic models of the hypersonic vehicle and complex constraints such as overload, dynamic pressure, and terminal impact angle are ignored, which require further study.

Data Availability

The simulation and experimental data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (61903379) and Research Project of State Key Laboratory of Laser Propulsion & Application (SKLLPA-14).

References

- [1] B. Xu and Z. K. Shi, "An overview on flight dynamics and control approaches for hypersonic vehicles," *Science China Information Sciences*, vol. 58, no. 7, pp. 1–19, 2015.
- [2] J. H. Wang, L. H. Liu, P. Wang, and G. J. Tang, "Integrated guidance and control for hypersonic vehicles in dive phase," *Acta Aeronautica et Astronautica Sinica*, vol. 38, p. 320328, 2017.
- [3] H. Yan, *Research on Robust Nonlinear Integrated Guidance and Control Design*, University of Science and Technology of China, Hefei, China, 2013.
- [4] T. Zhao, P. Wang, L. H. Liu, and J. Wu, "Integrated guidance and control with L2 disturbance attenuation for hypersonic vehicles," *Advances in Space Research*, vol. 57, no. 12, pp. 2519–2528, 2016.
- [5] B. Panchal, N. Mate, and S. E. Talole, "Continuous-time predictive control-based integrated guidance and control," *Journal of Guidance, Control, and Dynamics*, vol. 40, no. 7, pp. 1579–1595, 2017.
- [6] T. Yamasaki, S. N. Balakrishnan, and H. Takano, "Separate-channel integrated guidance and autopilot for automatic path-following," *Journal of Guidance, Control, and Dynamics*, vol. 36, no. 1, pp. 25–34, 2013.
- [7] C. Guo and X. G. Liang, "Integrated guidance and control based on block backstepping sliding mode and dynamic control allocation," *Proceedings of the Institution of Mechanical Engineers-Part G: Journal of Aerospace Engineering*, vol. 229, no. 9, pp. 1559–1574, 2015.
- [8] X. L. Liang, M. Z. Hou, and G. R. Duan, "Adaptive dynamic surface control for integrated missile guidance and autopilot in the presence of input saturation," *Journal of Aerospace Engineering*, vol. 28, no. 5, Article ID 04014121, 2015.
- [9] H. Zhou, H. Zhao, H. Q. Huang, and X. Zhao, "Integrated guidance and control design of the suicide UCAV for terminal attack," *Journal of Systems Engineering and Electronics*, vol. 28, no. 3, pp. 546–555, 2017.
- [10] X. D. Liu, W. W. Huang, and L. F. Du, "An integrated guidance and control approach in three-dimensional space for hypersonic missile constrained by impact angles," *ISA Transactions*, vol. 66, pp. 164–175, 2017.
- [11] W. H. Wang, S. F. Xiong, S. Wang, S. Y. Song, and C. Lai, "Three dimensional impact angle constrained integrated guidance and control for missiles with input saturation and actuator failure," *Aerospace Science and Technology*, vol. 53, pp. 169–187, 2016.
- [12] S. Wang, W. H. Wang, and S. F. Xiong, "Impact angle constrained three-dimensional integrated guidance and control for STT missile in the presence of input saturation," *ISA Transactions*, vol. 64, pp. 151–160, 2016.
- [13] M. Z. Hou and G. R. Duan, "Adaptive dynamic surface control for integrated missile guidance and autopilot," *International Journal of Automation and Computing*, vol. 8, no. 1, pp. 122–127, 2011.
- [14] H. T. Song and T. Zhang, "Fast robust integrated guidance and control design of interceptors," *IEEE Transactions on Control Systems Technology*, vol. 24, no. 1, pp. 349–356, 2016.
- [15] X. H. Wang and J. Z. Wang, "Partial integrated missile guidance and control with finite time convergence," *Journal of Guidance, Control, and Dynamics*, vol. 36, no. 5, pp. 1399–1409, 2013.
- [16] X. H. Wang and J. Z. Wang, "Partial integrated guidance and control for missiles with three-dimensional impact angle constraints," *Journal of Guidance, Control, and Dynamics*, vol. 37, no. 2, pp. 644–657, 2014.
- [17] R. Padhi, C. Chawla, and P. G. Das, "Partial integrated guidance and control of interceptors for high-speed ballistic targets," *Journal of Guidance, Control, and Dynamics*, vol. 37, no. 1, pp. 149–163, 2014.
- [18] X. L. Shao and H. L. Wang, "Back-stepping active disturbance rejection control design for integrated missile guidance and control system via reduced-order ESO," *ISA Transactions*, vol. 57, pp. 10–22, 2015.
- [19] C. Lai, W. H. Wang, Z. H. Liu, T. T. Liang, and S. A. Yan, "Three-dimensional impact angle constrained partial integrated guidance and control with finite-time convergence," *IEEE Access*, vol. 6, pp. 53833–53853, 2018.
- [20] X. H. Zhou, W. H. Wang, Z. H. Liu, C. Liang, and C. Lai, "Impact angle constrained three-dimensional integrated guidance and control based on fractional integral terminal sliding mode control," *IEEE Access*, vol. 7, pp. 126857–126870, 2019.
- [21] C. Ming, X. M. Wang, and R. S. Sun, "A novel non-singular terminal sliding mode control-based integrated missile guidance and control with impact angle constraint," *Aerospace Science and Technology*, vol. 94, Article ID 105368, 2019.
- [22] S. Jiang, F. Q. Tian, and S. Y. Sun, "Integrated guidance and control design of rolling-guided projectile based on adaptive fuzzy control with multiple constraints," *Mathematical Problems in Engineering*, vol. 2019, Article ID 6309462, 17 pages, 2019.
- [23] S. Jiang, F. Q. Tian, S. Y. Sun, and W. G. Liang, "Integrated guidance and control of guided projectile with multiple constraints based on fuzzy adaptive and dynamic surface," *Defence Technology*, vol. 16, no. 6, pp. 1130–1141, 2020.
- [24] G. L. Li, T. Chao, S. Y. Wang, and M. Yang, "Integrated guidance and control for the fixed-trim vehicle against the maneuvering target," *International Journal of Control, Automation and Systems*, vol. 18, no. 6, pp. 1518–1529, 2020.
- [25] J. Chang, Z. Y. Guo, J. Cieslak, and W. S. Chen, "Integrated guidance and control design for the hypersonic interceptor based on adaptive incremental backstepping technique," *Aerospace Science and Technology*, vol. 89, pp. 318–332, 2019.
- [26] J. Yi, D. F. Lin, P. Pei, X. W. Shi, and W. Wang, "Robust partial integrated guidance and control approaches for maneuvering targets," *International Journal of Robust and Nonlinear Control*, vol. 29, pp. 6522–6541, 2019.
- [27] X. L. Ai, Y. C. Shen, and L. L. Wang, "Adaptive integrated guidance and control for impact angle constrained interception with actuator saturation," *Aeronautical Journal*, vol. 123, no. 1267, pp. 1437–1453, 2019.
- [28] Z. Zhu, D. Xu, J. M. Liu, and Y. Q. Xia, "Missile guidance law based on extended state observer," *IEEE Transactions on Industrial Electronics*, vol. 60, no. 12, pp. 5882–5891, 2013.
- [29] S. Keshmiri, R. Colgren, and M. Mirmirani, "Six doF nonlinear equations of motion for a generic hypersonic vehicle," in *Proceedings of the 2007 AIAA Atmospheric Flight Mechanics Conference and Exhibit Hilton Head*, Hilton Head, South Carolina, August 2007.

Research Article

Analysis of Assembly Tolerance Based on Assembly Constraint Information Model

Chunxi Li  and Wenjun Hou

Human Computer Interaction Laboratory, Beijing University of Posts and Telecommunications, Beijing 100876, China

Correspondence should be addressed to Chunxi Li; lichunxi@bupt.edu.cn

Received 18 June 2021; Accepted 19 July 2021; Published 9 August 2021

Academic Editor: Jie Chen

Copyright © 2021 Chunxi Li and Wenjun Hou. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Mechanical products are composed of two or more parts. The geometric tolerance and dimensional tolerance of each feature in part will affect the assembly performance of the product, which are accumulated and propagated between assembly fit and parts. In this paper, through the secondary development of CAD software, the B-rep model of parts is obtained. The model information is decomposed and simplified based on geometric features to obtain the key information of parts in the assembly process, simplify the operation, and improve the accuracy. Through a directed graph network, the transmission model of assembly error information based on geometrical and dimensional tolerances (GD&T) on the surface of parts is established. Combined with the error transfer characteristics of different geometric surfaces and different error sources, guided by the breadth-first search algorithm and the shortest path theory, the search and establishment of a three-dimensional assembly chain are realized. Finally, the three-dimensional chain is simulated by the Monte Carlo method. The calculation results are compared with the error range obtained by the traditional method to prove the effectiveness of the method.

1. Introduction

Assembly is an integrative process of joining a complete mechanical product by bringing all individual parts together. There is a deviation between the nominal size and the actual size of the product. The manufacturing deviation and fitting deviation are propagated and superimposed in the component through the assembly of parts, affecting the assembly quality and expected function of products. Through tolerance analysis and tolerance stack-up, engineers can predict the final assembly accuracy before machining, so it has become a hot research area of automation [1]. Tolerance stacks are a direct and straightforward method to simulate dimensional deviation on the distance between different assembly features. It usually only includes dimensional tolerance (DT), although modern modification of this method also considers geometric tolerance [2]. However, under the actual manufacturing conditions, the propagation of geometric, dimensional tolerances (GD&T), the gap between mating parts, and the

assembly sequence of parts will cause some mating features to deviate from their nominal positions, thus affecting the assembly accuracy [3].

In recent years, scholars have established general tolerance analysis methods [4–8], which can obtain the assembly accuracy of mechanical products to a certain extent. However, few researchers consider the influence of the assembly sequence on assembly accuracy. Assembly sequence planning and assembly accuracy calculation depend on the assembly information model. To analyze the influence of the assembly sequence in the calculation of assembly accuracy, we should establish an appropriate model that can carry out assembly sequence planning and assembly accuracy calculation. The appropriate model needs to include the dimension and tolerance information of parts and the assembly relationship information between parts and components for assembly [4]. GD&T and assembly relationship information are added into parts' geometric features and propagated through parts' direct geometric contact.

To model assembly information, several problems need to be further studied. First of all, only some features influence the assembly process, so we are only aiming the key features to do the subsequent calculation. Secondly, the surface of parts determines geometric constraints between parts and geometric feasibility of the assembly. Non-geometric constraints have an influence on assembly efficiency. It is necessary to combine the nongeometric constraints that affect the assembly sequence to optimize the feasible geometric sequence. Different tolerances on different types of assembly surfaces will have different effects on assembly accuracy, so it is necessary to establish an assembly error transfer model to obtain error transfer and superposition in parts.

This paper presents a method to obtain the assembly dimension relation model. The key features of part obtained by simplifying the part and determining the influence area of error according to the different contact surfaces. Then, the complete assembly dimension relationship model is formed by combining the assembly error. Based on the model, the three-dimension chain between the corresponding features of two different parts of the assembly is searched by the graph search algorithm and solving the assembly parts' assembly error. The basic process is as follows: (1) the hierarchical assembly information model can establish the decomposition of the geometric feature of the part represented by the B-rep model. (2) The assembly information model carries out assembly sequence planning. (3) On the premise of obtaining the optimal assembly sequence under multiple constraints, the three-dimensional chain of parts in an assembly accuracy propagation model is searched. (4) The assembly accuracy of related parts is calculated based on obtaining the dimension chain.

2. Literature Review

2.1. Assembly Constraint Model. The optimal assembly sequence has an important influence on assembly time and cost [9, 10]. For generating the valid assembly sequences, the engineers must be pointed out the assembly precedence between parts. However, it is challenging and time-consuming to find the optimal assembly sequence due to many feasible assembly sequences. Assembly sequence planning (ASP) is typical of the discrete optimization problem in mathematics. In the early 1960s, researchers developed at least one optimal feasible assembly sequence for a product [11]. Then, through the assembly constraints between parts, the feasible assembly sequence can be found using the graphical method [12]. Although these traditional methods provide a correct and complete solution for the assembly configuration of some parts counting, they were time-consuming and required many calculations [13]. Also, the methods were semi-automated, which required skilled users to oversee the decision-making process.

In 1988, graphic representation methods such as connection graph, interference graph, and assembly constraint graph were proposed to represent assembly trajectory constraints and were used to generate assembly priority relationships [14]. de Mello and Sanderson introduced all

feasible assembly sequences of the product with the and/or graph [15]. In the graph, the node is an assembly subpart, the root node is a complete part, the leaf is a separate part, and the arc corresponds to a feasible assembly operation, linking each node to all combinations in which it can be split [16]. Wolter, on the other hand, used a partial assembly tree to express the assembly. Its nodes correspond to subassemblies in the assembly process. The leaf node is a single part, while the root node is the final mechanical product [17]. Besides, there is an assembly directed graph [9, 18], whose nodes correspond to the assembly of parts in the assembly and whose arcs correspond to feasible assembly operations [19].

A researcher prefers extracting assembly sequence constraints from Computer-Aided Design (CAD) software to minimize human intervention. Some researchers have developed automatic extraction methods to retrieve assembly connection data, interference data, and assembly stability information through CAD interfaces [20, 21]. However, this method requires manual operations, which rely on engineers' experience, so it is not the optimal solution as human input errors are a significant factor [22]. The existing methods cannot automatically establish the appropriate assembly constraint information model. In this paper, the surface included in the part is divided to obtain the geometric structure and the topological relationship between the geometric structures. The introduction of model simplification reduces the number of nodes in the model, improves operation efficiency, and results in faster and more accurate results. The key features that affect the assembly accuracy are combined with the nongeometric information to form a complete assembly information model.

2.2. Error Transformation Model for Assembly. The design and calculation of dimensions and tolerances are difficult problems in the industry. Scholars have carried out much research on the mathematical model expression of assembly tolerance to solve this problem. Typical tolerance models include topologically and technologically related surface (TTRS) [7, 23], polychromatic set model [24], small-displacement sensor (SDT) [25], tolerance map (T-map) [5], and degree of freedom model [26].

As a widely used model, Clément et al. [27] proposed the TTRS model on seven-element surfaces, and twenty-eight surface relationships determine the tolerance. Davidson and Shah [28] established the T-map tolerance model of standard plane features. On this basis, Xiao and Zhu [29] combined with T-map and ASME standard, took the intersection axis as the research object, established the spatial domain of assembly feature deviation fluctuation, and improved the accuracy of the transformed tolerance. Solving the problem that the machining datum does not coincide with the design datum in the machining process, the tolerance graph is another model for calculating tolerance in the n -dimensional Euclidean space, mapping the geometric features of parts to the point space model. Ke et al. [30] established an M-map to realize the transfer of the datum plane and realized the adjustment and optimization of machining tolerance under different constraint conditions.

The stack-up of assembly errors is another problem in the process of assembly accuracy calculation. Liu et al. [31] divided the assembly errors into three categories: geometric position deviation, geometric shape deviation, and assembly position deviation, and then, they established the assembly directed graph to express the transmission process of assembly accuracy. Boccaletti et al. [32] divided a complex network into different modules based on complex network and dynamic cluster coefficient analysis, which provides a new idea of error tracing. Based on the above principles [33], Zhu et al. [34] established the deviation transfer network model of the mechanical assembly process, identified the key assembly surface, and realized the error traceability. The existing methods are limited to the transfer direction of assembly error. In addition, few researchers regard the assembly sequence as one of the constraints of assembly accuracy. However, different assembly sequences will lead to different contact sequences between parts and different error transmission paths. In this paper, in the same model, the optimal assembly sequence of parts is deduced, and the assembly accuracy of parts is obtained under this sequence.

2.3. Problem Definition and Objectives of Present Research. Every feature in a component is subjected to variation between design positions and is strictly applied with some tolerances. When there are multiple components in a mechanical product, the variations will accumulate and propagate, affecting the functional performance of parts. The sources of these changes can be divided into three categories:

- (1) The size change or tolerance superposition of a single part
- (2) The change of geometric features
- (3) The change of parts and the change of transmission between parts in the assembly process

In previous studies, the second and third types of deviations are not considered in evaluating fit accuracy, which usually regards as null according to the translation or selection changes added in the assembly process.

The variation of geometric features can significantly affect tolerance superposition, depending on the type of translation and rotation variation added during assembly. In the process of assembly accuracy simulation, considering the assembly sequence of mechanical products and the matching accuracy between parts, the assembly accuracy obtained is more realistic than the simple dimensional accuracy. The propagation direction of different error sources on different surfaces is also different. In searching the dimension chain, it is necessary to determine whether each error has a component in the calculation direction of assembly accuracy and then ignore the interaction of multiple errors. So, the expected accuracy of precision assembly can be calculated, and the service life of parts can be improved.

This paper aims to apply geometric tolerance and dimensional tolerance, surface tolerance, assembly error, and assembly sequence of parts to the calculation of assembly accuracy. Therefore, it is necessary to establish a model containing all the assembly information of the parts. The

influence of geometric constraints between parts on the feasible motion range of parts is analyzed. It is important to determine the feasible transfer direction of assembly deviation of different kinds of error sources on different surfaces. In establishing the model, the key features of the parts participating in the assembly process are obtained through simplification and planning the assembly sequence of the products. After obtaining the contact sequence between parts, it is necessary to analyze the influence of GD&T deviation on assembly accuracy. We propose a breadth-first algorithm combined with the assembly sequence to search the 3D chain of the required assembly size in the assembly error transfer model to solve this problem. Finally, the assembly accuracy of the product is obtained by the Monte Carlo method. The shaft and its related parts of the two-stage reducer are an example. Analyzing is carried out in order to understand the proposed method.

3. Assembly Information Model

The assembly constraint relation model expresses constraint and mating relations between parts. Assembly tolerances can be used as attributes of assembly models and are attached to them. This paper builds a hierarchical representation model whose structure is shown in Figure 1. The representation model includes five levels: assembly level, part level, assembly feature level, surface level, and assembly tolerance level.

To meet the design requirements, CAD models usually contain accurate geometric information. There are some structures in part, such as chamfering, fillet, or tool retracting groove, which will increase the number of nodes in the assembly information model and increase calculation difficulty. Therefore, it is necessary to calculate the key features of parts to improve operational efficiency. In this paper, by simplifying the model extracted from CAD, the key features with assembly constraints are obtained. If a simplified basic unit is a surface, the part may not be a complete entity. To avoid this problem, take the feature as the minimum simplified element. Because different engineers understand that the CAD modeling process is different, using feature trees directly in CAD software cannot get the simplest simplified results. Therefore, through the secondary development of the CAD system, the digital model of part B-rep representation is obtained. This method generates a feature-based model from the B-rep model by applying the volume decomposition method. The assembly information model of mechanical products is constructed hierarchically to store the assembly constraints and the nongeometric information related to the parts.

3.1. B-Rep Model. Obtaining the geometric constraint information between parts, through the contact surface between parts, we can obtain the geometric structure of parts, including contact surface and noncontact surface. The noncontact surface ensures the transmission of geometric constraints on the part, while the contact surface ensures the transmission of geometric constraints between the part. The

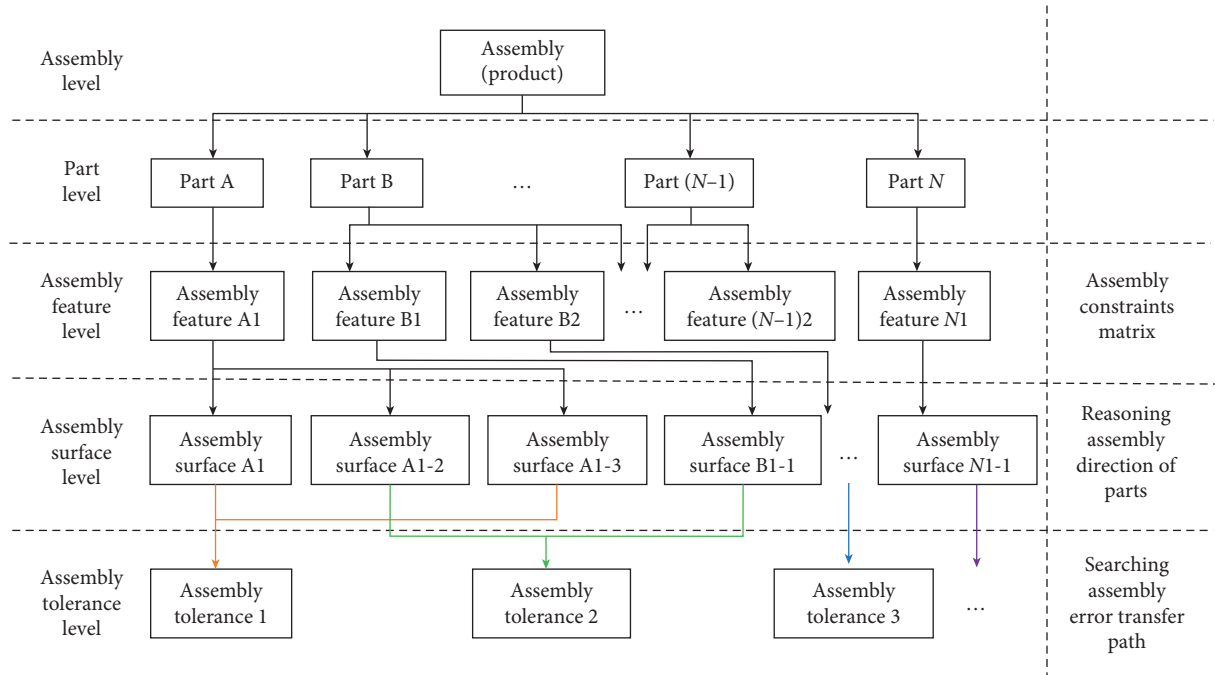


FIGURE 1: Assembly information model of the product.

B-rep model established the model of parts' contact relationship through the secondary development of CAD. The feature tree of the part is obtained through the decomposition of the volume feature.

The assembly information model can represent constraint information at each level and indicate the reasoning relation matrices built at different levels. Boundary representation is a common geometric description method in CAD. The B-rep model retains the information of points, lines, and surfaces in the CAD model, which helps calculate the geometric features of parts. In the B-rep model, as shown in Figure 2, the type of part surface is obtained, and the concave-convex of the edge formed by the intersection of two surfaces is judged by the angle between the normal vectors of two surfaces.

3.2. Geometric Feature Decomposition. The mechanical product can be seen as a complex geometry formed by many simple subfeatures and their Boolean operation. When applying additive features, convex inner loops are often generated at feature intersections. So, the convex inner loops can be the sign of additive features. Koo and Lee [35] proposed the wrap-around operation to decompose parts. The algorithm for the wrap-around operation on a given shape S is described as follows:

- (1) Find and mark the convex inner loops $\{L_i\}$ of S (Figure 3(a)).
- (2) Separate every faces $\{F_j\}$ (Figure 3(b)).
- (3) Find $\{L_i\}$ in $\{F_j\}$ (Figure 3(b)).
- (4) Remove $\{L_i\}$ and add new surface according to $\{L_i\}$ (Figure 3(c)).

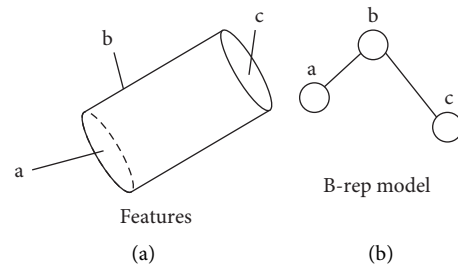


FIGURE 2: B-rep model for parts.

- (5) Sew all $\{F_j\}$ and add corresponding plane (Figure 3(d)).
- (6) Unite T and W with a regularized Boolean union (Figure 3(d)).
- (7) If T has convex inner loops, apply the wrap-around operation again to T_1 and obtain T_2 . Repeat the wrap-around operation until convex inner loops do not exist.

3.3. Assembly Constraints Model. Most studies on the simplification of feature-based 3D CAD models involve the following two steps: feature rearrangement according to importance and removal of features with low importance until the termination condition. In this study, through the effective volume of features and the contact features between parts, the parts are simplified on the premise of maintaining the connectivity of the model using a feature adjacency graph.

A feature tree is generated by decomposing the part and reconstructing the simple subfeatures. Different deletion

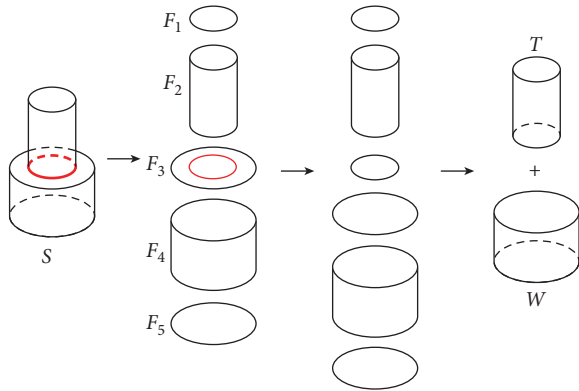


FIGURE 3: Concept of volume decomposition.

processes lead to different results, so it is necessary to determine a reasonable deletion order to ensure the integrity of the constraint information of the final part. Kim and Mun [36] and Kang et al. [37] proposed several recombination strategies. This study evaluates the importance of subfeatures according to the following criteria. The main results are as follows:

- (1) The importance of additive features is higher than that of subtractive features.
- (2) Larger features are more important than smaller ones.
- (3) Features need to be retained in contact with other parts.
- (4) The first adjacent feature near the port is more important than the other features. We use the following formula to measure the importance of features:

$$f = n \times (C_1 + C_2 + C_3). \quad (1)$$

If it is a port feature, it cannot be deleted and retained. At this time, $n = 1$, making $f > 0$. Otherwise, $n = -1$ making $f < 0$. C_1 is the volume ratio of the feature to the largest feature in the part, which is a number greater than 0 and less than 1. If it is an additive feature, $C_2 = 1$; otherwise, it is 0. Suppose that it is the first feature near the port, $C_3 = 1$. Otherwise, it is 0. The results are shown in Figure 4.

As shown in Figure 5, all the nodes are connected, and all the nodes in the graph are ordinary nodes. When deleting node 4, node 2 will become the cut node, while node 3, node 5, and node 1 remain unchanged. After the deletion of node 3 on the above condition, all the nodes in the graph become ordinary nodes. Thus, the graph nodes can switch between cut nodes and ordinary nodes in the simplification process. Furthermore, the set of cut nodes need to be reevaluated when the graph changes.

During the reduction of nodes on the diagram in the graph above, the nature of the retained node as a cut node also changes. Therefore, it is necessary to determine whether the node in the graph is a cut node after deleting a feature. When the node to be deleted is a cut point in the geometric contact diagram, skip the feature. Then, the part model is simplified according to the order.

The assembly constraints graph is highlighting the features from the same part in the same color. The node code is a 4-digit string. The first two digits are the *ID* of the part, and the last two digits are the *ID* of the face in part, as shown in Figure 6(d).

3.4. Assembly Variation Model. A complex component is considered to have a high number of parts and a large degree of connectivity between subfeatures or parts. Most of the existing assembly accuracy analysis methods simplify the machining error of the part surface to the translation and rotation of the ideal geometric surface along with its theoretical position. This simplification ignores the influence of surface morphology on the position of parts, thus affecting the accuracy of precision analysis results.

The assembly tolerances mentioned in this paper include dimensional tolerances, shape tolerances, and position tolerances. Form tolerances and position tolerances are collectively referred to as geometric tolerances. The error of the part can be regarded as an ideal surface, which is translated and rotated under the limitation of orientation and positioning error and then superimposed with the deformation caused by shape tolerance. Therefore, the small-displacement torsor (SDT) represents the orientation and positioning error of the part, and then, the shape error is expressed in the form of the point cloud. The superposition of the two is the final assembly error .

(a) *Positioning Tolerance Modeling.*

When geometric features are constrained by location tolerance and orientation tolerance, location and orientation errors can be expressed by SDT, which include a set of translation components along the x -axis, y -axis, and z -axis ($[u, v, w]$) and a set of rotation components along the x -axis, y -axis, and z -axis ($[\alpha, \beta, \gamma]$). The SDT representation of common geometric features and their value range is shown in Table 1. In Table 1, L and W represent the length and width of the rectangular plane, φd and φD represent the outer diameter of the circle, and H and h represent the height of the cylinder and cone. On this basis [38], the geometric surface with positioning/orientation errors can be obtained by translating and rotating the geometric surface in the ideal position, and the given SDT parameters determine the values of translation and rotation.

(b) *Form Tolerance Modeling.*

Similar to positioning or orientation tolerance, form tolerance also limits the variation range of geometric features. The rectangular plane is taken as an example; its form tolerance region is also the area between two parallel planes, as shown in Figure 7. The nonideal surface generated by form error modeling still needs to meet the constraints of form tolerance. Form error can be modeled by point cloud simulation.

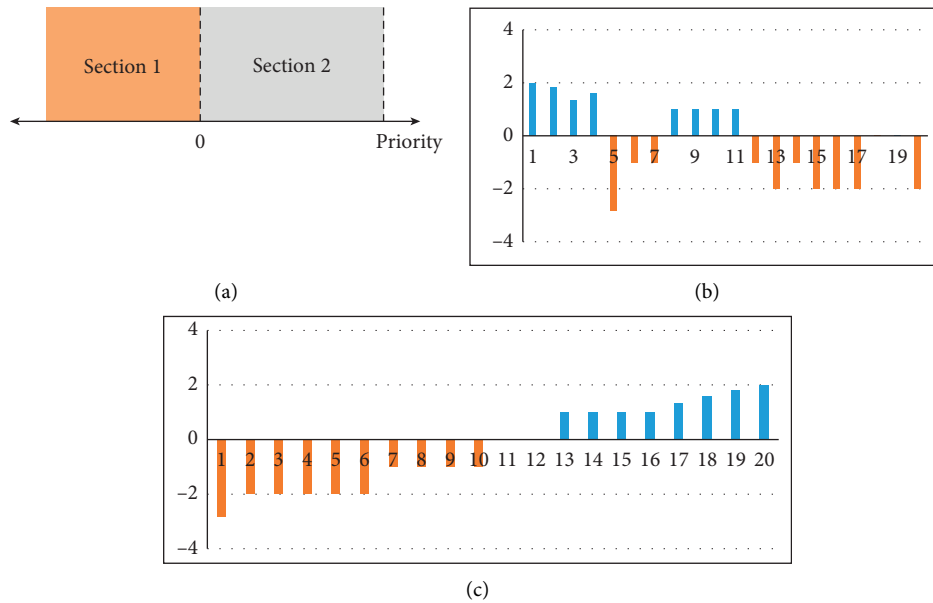


FIGURE 4: Geometric features are rearranged according to their importance. (a) Two sections according to feature importance. (b) Importance of the feature. (c) Rearrange the features according to their importance.

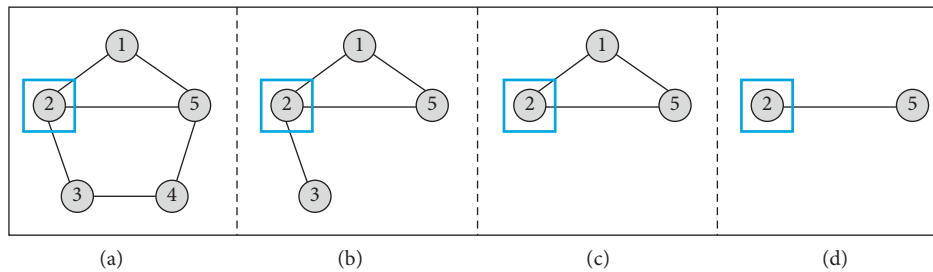


FIGURE 5: The change of cut points in the process of node deletion in the undirected graph.

The nonideal surface model is generated by superimposing a randomly generated geometric surface with form error on the geometric surface with location and orientation error based on the modeling of location, orientation error, and form error, as shown in Figure 8. Form tolerance and position tolerance are independent of each other, and they constrain the surface of parts together. It should be noted that, in the process of using numerical simulation, in some cases, the plane will exceed the constraint range of tolerance, so manual correction is needed.

4. Assembly Sequence Planning

The essence of part assembly is to impose constraints on geometric elements to adjust the position and the rotation of the part model. The geometric elements involved in the constraints between parts include various geometric surfaces, such as planes, cylinders, and cones. This paper focuses on the geometric surfaces with contact relationships between two different parts.

4.1. Geometrically Feasible Assembly Sequence. In 3D space, the assembly constraints of different kinds of geometric surfaces are transformed into the remaining feasible motion direction of the part. The unit spherical coordinates are stored in the form of a set, as shown in Figure 9.

Table 2 shows the constraints of the geometric contact surface in part on the feasible direction of movement and the assembly direction. When there are constraints between the planes, the part can move in the opposite hemisphere of the plane normal vector, and the assembly direction is the normal direction of the plane; when there are cylindrical constraints, the part can move in the positive and negative direction of the vector, and the assembly direction is also the positive and negative direction of the vector; when there are conical constraints, the part can move in the opposite direction of the vector, and the assembly direction is the component of the vector.

The potential assembly sequence for N parts in a mechanical product is $N!$. Geometric constraints between parts result in a reduction in the number of parts. The method of

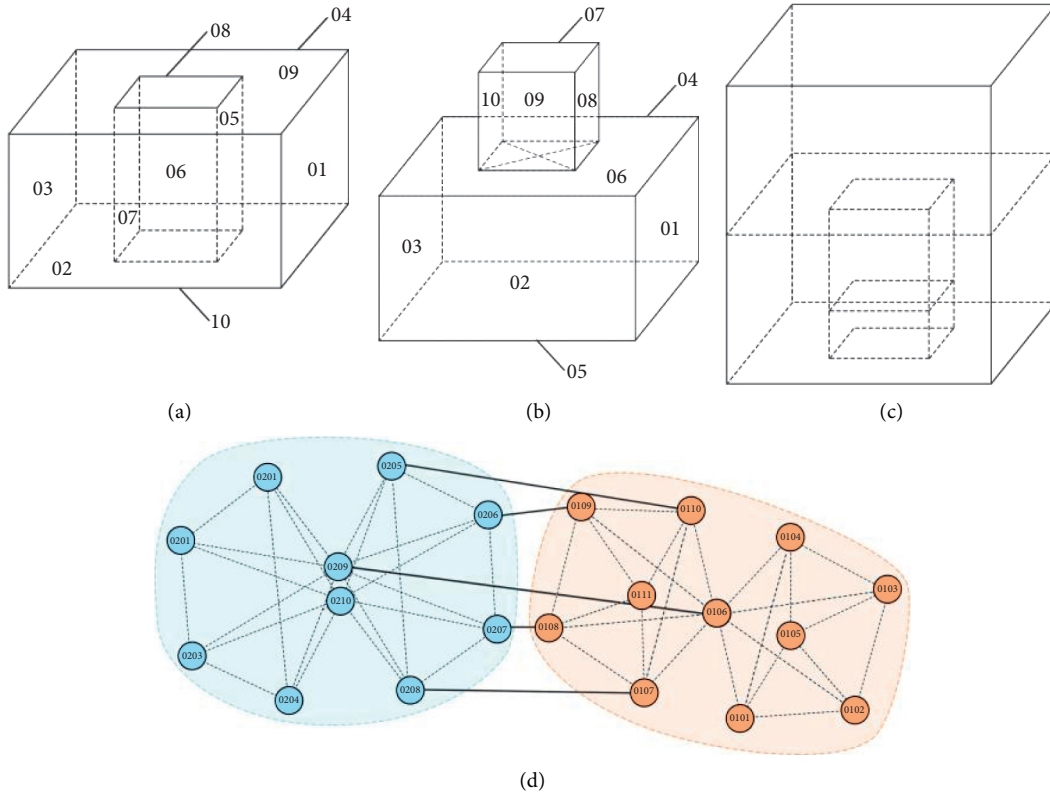


FIGURE 6: Contact relationship of parts based on surface representation. (a) Part 1 CAD model. (b) Part 2 CAD model. (c) Fitting model of parts. (d) Feature representation of part fitting model.

TABLE 1: Analysis of the tolerance zone in common geometric features.

	Plane	Cylinder	Cone
Tolerance zone			
Deviation vector	$(0, 0, \delta_z, \delta_\alpha, \delta_\beta, 0)$	$(\delta_x, \delta_y, 0, \delta_\alpha, \delta_\beta, 0)$	$(\delta_x, \delta_y, \delta_z, \delta_\alpha, \delta_\beta, 0)$

obtaining the assembly sequence is to inverse the disassembly sequence. In this paper, a geometrically feasible assembly sequence for all parts is obtained by continuously removing the parts from the product.

An undirected graph of assembly constraints of mechanical products is established through the port features of parts and the B-rep model of assembly. The node represents the contact surface between parts and records the type and direction of the surface. The assembly sequence is represented as a string, so the sequence [1 3 4 5 6 7 8 9 10 15 16 2 12 11 13] is represented as “01, 03, 04, 05, 06, 07, 08, 09, 10, 15, 16, 02, 12, 11, 13,” and each part number is represented by

two digits. Part 2 has ID number 02, and part 13 has ID number 13.

Parts with the not empty feasible moving area can be disassembled when constraints are received. Then, surface nodes and connections between surface nodes and link edges can be deleted, which contact other parts. Record the removed parts, update the assembly constraint relationship model, and then, obtain the parts with no empty feasible moving area again. When there are several removable parts, store them separately to form several strings representing the disassembly sequence. Repeat this process until all parts are completely removed. Invert the obtained disassembly

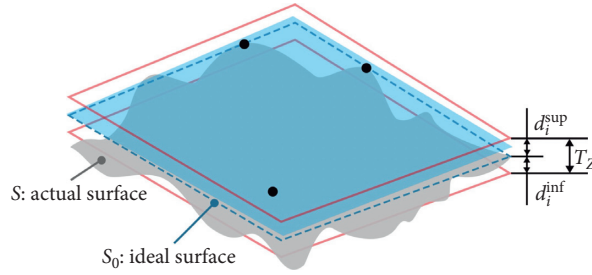


FIGURE 7: Form error of the plane.

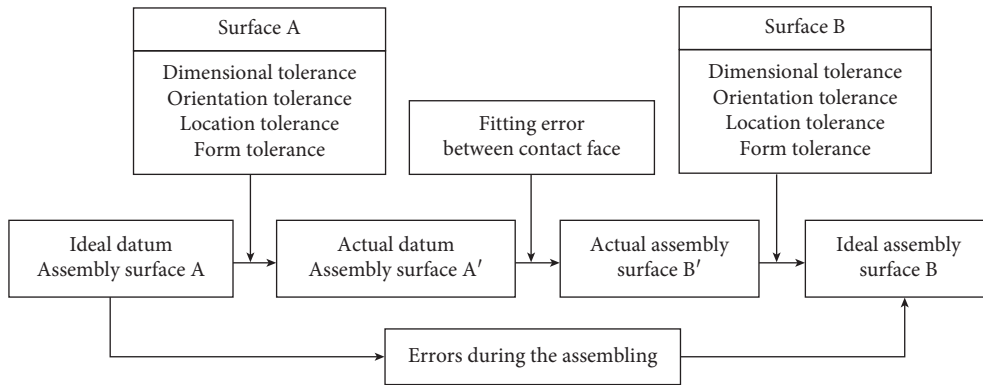


FIGURE 8: Error transfer between two surfaces.

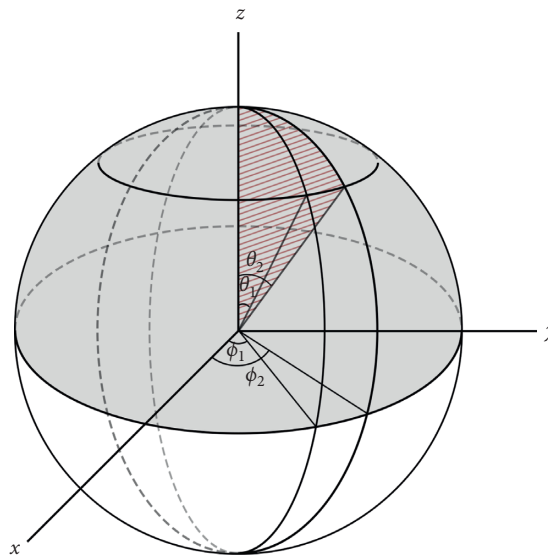


FIGURE 9: Feasible assembly region represented by spherical coordinates.

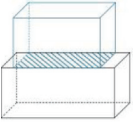
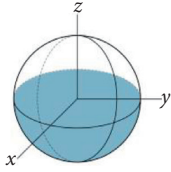
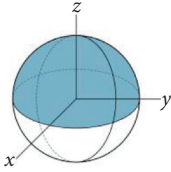
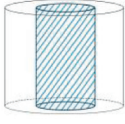
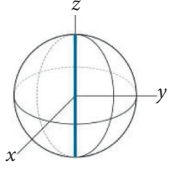
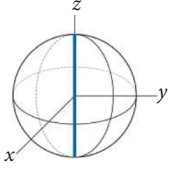

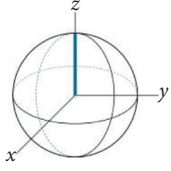
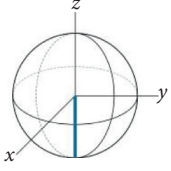
sequence to define the geometrically feasible assembly sequence of a part.

For example, there are two pairs of surface contact in model A and model B. They are marked as $x_1, y_1, x_2,$ and $y_2,$ and the assembly direction constraint of the part between x_1 and x_2 is shown in Figure 10(a), and the constraint of the part between y_1 and y_2 is shown in Figure 10(b). The feasible

assembly domain of the two parts is shown in Figure 10(c). The details are shown in Figure 10.

4.2. Assembly Sequence Optimization. Before optimizing the assembly sequence, it is necessary to determine the fitness function, which describes the difficulty and complexity of

TABLE 2: Constraints of the surface on the feasible assembly region of parts.

Connect relationship	Assembly direction	Disassembly direction	Assembly zenith angle	Assembly azimuth angle	Disassembly zenith angle	Disassembly azimuth angle
			$\theta = [(\pi/2), \pi]$	$\phi = [0, 2\pi]$	$\theta = [0, (\pi/2)]$	$\phi = [0, 2\pi]$
			$\theta = [0, 0]$	$\phi = [0, 2\pi]$	$\theta = [0, 0]$	$\phi = [0, 2\pi]$
			$\theta = [\pi, \pi]$	$\phi = [0, 2\pi]$	$\theta = [\pi, \pi]$	$\phi = [0, 2\pi]$

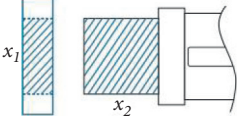
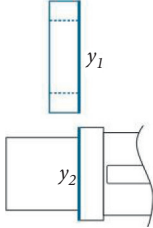
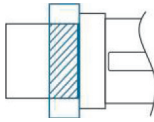
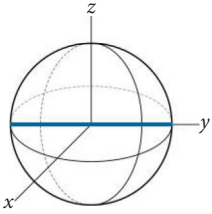
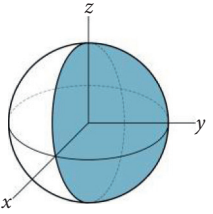
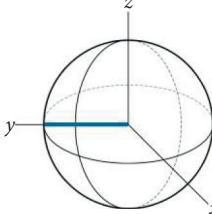
	(a)	(b)	(c)
Assembly constraint			
Feasible assembly direction			

FIGURE 10: Determination of the assembly region between parts.

assembly sequence operation. The fitness function includes the following:

- (1) The number of tools changes in the assembly process
- (2) The number of times the assembly direction changes in the assembly process
- (3) The fitness difficulty between parts

The optimization function used in this paper is shown in formula (2), and the calculation method of variables is shown in Table 3.

F_n represents the assembly process factors, including the change times of assembly tools R_1 , the difficulty of assembly operation R_2 , and the change times of assembly direction R_3 . Xie and Zhong used the analytic hierarchy process to evaluate the assembly sequence, in which the weights of various constraints were obtained by consulting experts using the radi scale. This paper referred to the relevant conclusions [39]. The value of weight factors was set as $\omega_1=0.2$, $\omega_2=0.3$, and $\omega_3=0.5$. Then, using the fitness function to select the geometric feasible assembly sequence to get the optimal solution under the constraint:

TABLE 3: Calculation formula of the fitness function.

The number of times the assembly tool changes	Assembly operation difficulty	Number of assembly direction changes
$\text{Tool_Change}_i = \begin{cases} 0 & \text{if tool_type}_i = \text{tool_type}_{i-1} \\ 1 & \text{otherwise} \end{cases}$	$\text{Ass_Diff} = \begin{cases} 0 & \text{clearance fit} \\ F_i/F_{\max} & \text{transition fit} \\ 1 & \text{interference fit} \end{cases}$	$\text{Ass_dir_changes} = \begin{cases} 0 & \theta_i \cap \theta \neq \emptyset \&\&\varphi_i \cap \varphi \neq \emptyset \\ 1 & \text{otherwise} \end{cases}$
$\text{Num_Tool_Change}_i = \sum_{j=1}^{i-1} \text{Tool_Change}_j$	$\text{Ass_Force} = \sum_j^i \text{Ass_Diff}_j$	$\text{Num_Dir_change}_i = \sum_{j=1}^{i-1} \text{Ass_dir_change}_j$
$R_1(i) = (\text{Num_Tool_Change}_i / \text{Num_Tool_Change}_{\max})$	$R_2(i) = (\text{Ass_Force}_i / \text{Ass_Force}_{\max})$	$R_3(i) = (\text{Num_Dir_Change}_i / \text{Num_Dir_Change}_{\max})$

$$F_n = \omega_1 R_1 + \omega_2 R_2 + \omega_3 R_3. \quad (2)$$

5. 3D Dimension Chain Search Based on the Breadth-First Algorithm

According to the assembly error model established in the previous paper, the assembly error of parts can be obtained by calculating the three-dimensional dimension chain of mechanical products. According to the assembly sequence, we need to obtain the error transfer path in each part of the product and the error transfer relationship between parts and gradually obtain the three-dimensional chain of this precision. When the number of nodes increases, the accumulated error on the path also increases, so transforming the assembly accuracy calculation problem can be transformed into a shortest path problem.

Based on the model network established in the previous paper, a directed graph G is constructed, in which the faces are mapped to the node n of G and the transfer relationship of assembly errors between faces is mapped to the edge e of G . For the edge e , if the precision level between two sides is high, the distance $W(e)$ of edge e is smaller. If the precision level between two sides is low, the distance $W(e)$ of edge e is larger. If there is no component of the error precision between two sides in the band direction, then $W(e)$ is a large number M . Let $d(v_i, v_j)$ be the set of paths with u and v as the endpoints in G and $W(d(v_i, v_j))$ denote the sum of the distances on the top of the path $d(v_i, v_j)$. The three-dimensional dimension chain search problem can be expressed as finding a path $P_0(v_i, v_j)$ in $d(v_i, v_j)$ in G such that

$$W(P_0(v_i, v_j)) = \min(W(p(u_i, v_j)) \in d(u_i, v_j)). \quad (3)$$

Typical shortest path algorithms include Dijkstra algorithm, Floyd algorithm, Bellman-Ford algorithm, and SPFA algorithm. In this paper, we get the shortest path when the starting vertex and the ending vertex are known. Therefore, we are choosing the Floyd algorithm as the search algorithm for the 3D dimension chain.

The Floyd algorithm thought is that the shortest path from any node v_i to any node v_j is no more than two possibilities, one is directly from node v_i to node v_j and the other is from v_i through several nodes to v_j . Suppose that distance between node u and node v , which is expressed as $\text{dist}(v_i, v_j)$, is the shortest path distance from node v_i to node v_j for each node k , and check whether $\text{dist}(v_i, k) + \text{dist}(k, v_j) < \text{dist}(v_i, v_j)$ holds. If it is true, it is proved that the path from node v_i to node k and then to v_j is shorter than the path from u to v . After traversing all nodes k , $\text{dist}(v_i, v_j)$ records the shortest path distance from v_i to v_j .

The error transfer graph $G = \{V, E\}$ of each part and the assembly sequence $\mathbf{s} = [a(i)]_n$ under multiple constraints are calculated. The weight adjacency matrix $\tau = [a(i, j)]_{n \times n}$ is constructed, and a subsequent node matrix path $\mathbf{p} = [a(i, j)]_{n \times n}$ is introduced to record the shortest path between two points. The specific algorithm steps are as follows:

- (1) Determine the transfer direction of the error. According to the start and end surfaces of assembly error, spherical coordinates are recording the error transfer direction.
- (2) This can be obtained between the contact surfaces and the contact order of the parts by the determined assembly sequence. Judge whether there is a component in the direction of error transmission. The error transfer path between parts, the participating parts, and the start and end surfaces of the participating parts is obtained
- (3) For each part, the weighted error transfer digraph is established from any path. The distance between all the two points is the weight of the edge. If there is no edge connection between the two points, the weight is infinite. The graph weight adjacency matrix and node matrix path are constructed.
- (4) Use the Floyd algorithm to find intermediate node k . If there is a vertex k , update the result.
- (5) Repeat the above step (2)–step (4) until all nodes are traversed until the end. The shortest path formed is the three-dimensional dimension chain of parts.
- (6) Through the dimension chain of parts and the fitting error between parts, the three-dimensional dimension chain of the error to be solved is formed.

6. Assembly Accuracy Calculation

After obtaining the three-dimensional dimension chain of the product, the assembly accuracy of the product can be calculated, including the error transfer and numerical calculation.

6.1. Error Transfer of Homogeneous Transformation Matrix. Based on the coordinate transformation theory of robot kinematics, a 4×4 homogeneous transformation matrix is used to express the tolerance model transmit [40]:

$$\mathbf{M} = \begin{bmatrix} 1 & -\gamma & \beta & u \\ \gamma & 1 & -\alpha & v \\ -\beta & \alpha & 1 & w \\ 0 & 0 & 0 & 1' \end{bmatrix}. \quad (4)$$

Then, the transformation relation of the coordinate position of P point from S_2 to S_1 is as follows:

$$\mathbf{P}_{s_2} = \begin{bmatrix} X_{s_2} \\ Y_{s_2} \\ Z_{s_2} \end{bmatrix} = \mathbf{T} \bullet \begin{bmatrix} X_{s_1} \\ Y_{s_1} \\ Z_{s_1} \end{bmatrix} = \mathbf{T} \bullet \mathbf{P}_{s_1}. \quad (5)$$

As shown in Figure 11, the nominal coordinate system and the actual coordinate system of the geometric features of the mating surface are established at the mating surface between parts. They represent mapping of the measuring point P from one space coordinate system to another space coordinate system by the homogeneous coordinate transformation matrix.

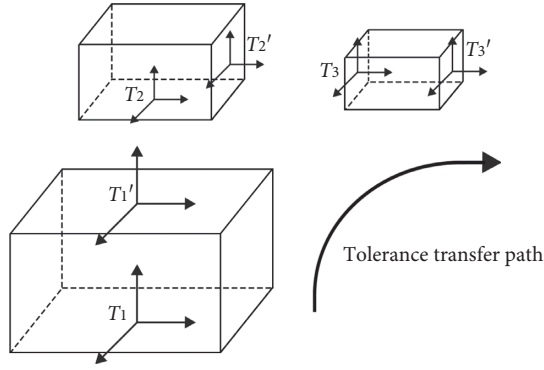


FIGURE 11: Tolerance transfer path in the product.

The transformation of any mating surface from the nominal feature coordinate system to the actual feature coordinate system will affect the spatial attitude of subsequent parts in the global coordinate system. For the measurement point P , its spatial position and pose in the global coordinate system will be affected by the surface changes of each link in the dimensional chain. It is mathematically described as a cross product of the series transformation matrices of all adjacent coordinate systems on the transfer path:

$$\mathbf{P}_{s3} = (\mathbf{T}_{11} \bullet \mathbf{T}_{12} \bullet \mathbf{T}_{22} \bullet \mathbf{T}_{23} \bullet \mathbf{T}_{33}) \mathbf{P}_{s1}. \quad (6)$$

Among them, \mathbf{T}_{11} , \mathbf{T}_{22} , and \mathbf{T}_{33} represent the coordinate transformation matrix of the actual mating surface relative to the nominal mating surface between parts \mathbf{P}_1 , \mathbf{P}_2 , and \mathbf{P}_3 . \mathbf{T}_{12} and \mathbf{T}_{23} represent the relative transformation matrix of nominal coordinates between parts \mathbf{P}_1 to \mathbf{P}_2 and \mathbf{P}_2 to \mathbf{P}_3 .

6.2. Calculation Accuracy Based on Monte Carlo Method.

In the process of calculating assembly accuracy, the process of tolerance superposition optimization is extremely complex. An optimization strategy combined with the Monte Carlo method has been proposed to solve this problem. According to the distribution law of the uncertain values in the error domain, the specific size of each assembly is randomly selected, and the error domain of the final assembly error is obtained by multiple stack calculation.

In the three-dimensional dimension chain of error transfer, for the numerical simulation of the errors contained in the nodes, the manufacturing errors usually conform to the normal distribution. Through the statistical simulation, the approximate solution of the assembly error is obtained. After determining the three-dimensional dimension chain of assembly error transfer, the number of simulation N needs to be determined, and the distribution function of error sources determines each error source in the chain, and random sampling is carried out within the error range. In the analysis of geometric tolerance, the parameters of SDT which control geometric features are randomly selected in their value range to obtain the corresponding random sequence. According to the stacking process shown in Figure 8, calculate and repeat this step until the number of repetitions meets the simulation number N . The maximum

and minimum values of the calculation results are the prediction range of the corresponding assembly accuracy.

7. Case Study

In this paper, the assembly of a shaft in the speed reducer is taken as an example for algorithm verification, implemented by Matlab 2017b and Solidworks API programming. The shaft is mainly composed of shaft, key, gear, locating ring, end cover, sleeves, two bearing, gaskets, and screw. In order to save analysis time and cost and improve disassembly efficiency, we use one fastener to express multiple fasteners. A three-dimensional model of a shaft in the speed reducer is shown in Figure 12 and Table 4. The fit relationship between parts in the assembly is shown in Table 5, and the rest fit relationship is clearance fit.

The model is decomposed into Boolean operation sets of several basic subfeatures according to the volume feature decomposition method. The types of planes included in the simple subfeatures are counted, the normal vectors and contact relationships of the planes are determined, and the assembly information model is established. According to the simplified method proposed in this paper, the port features of each part are determined and the importance of neutron characteristics of each part is ranked. The importance of each subfeature of the most complex shaft part is shown in Figure 4(c) and is then sequenced on this basis. Obtain the contact relationship between the features, and establish the undirected diagram of the contact relationship between the features, as shown in Table 6. Parts are gradually simplified by an algorithm until further simplification is not possible. The feasible movement direction of each part under assembly constraints and the geometrically feasible sequence of mechanical products are obtained through the contact relationship between the surfaces of the parts.

According to the results, the simplified shafts are composed of 9 features, and 11 are reduced. It retains all features of contact with other parts; the volume is 48.98 dm^3 before simplifying, the reduction volume is 37.55 dm^3 , the volume change is 11.86 dm^3 , and the reduction ratio is about 23.3%. It retains the number of surfaces of 46 before simplifying, and the number of surfaces after simplifying is 26, the number of surfaces reduced is 20, and the reduction percentage is about 43.5%. The characteristic topology of the simplified model has a change of 24.3%, where the node represents the feature in the part and the line represents the contact relationship between the two features. By simplifying the parts in the assembly and combining the contact relation between the parts, the following assembly relation characteristic diagram is obtained (Figure 13).

Based on the above analysis of geometric constraints between parts in the product, we obtain 36 geometrically feasible assembly sequences as the input of subsequent assembly sequence optimization, which refer to Table 7 for complete information.

Firstly, the disassembly sequence is reversed to obtain the geometrically feasible assembly sequence of the part, and the feasible direction of the part is determined according to the contact relationship of the direct surface of the part.

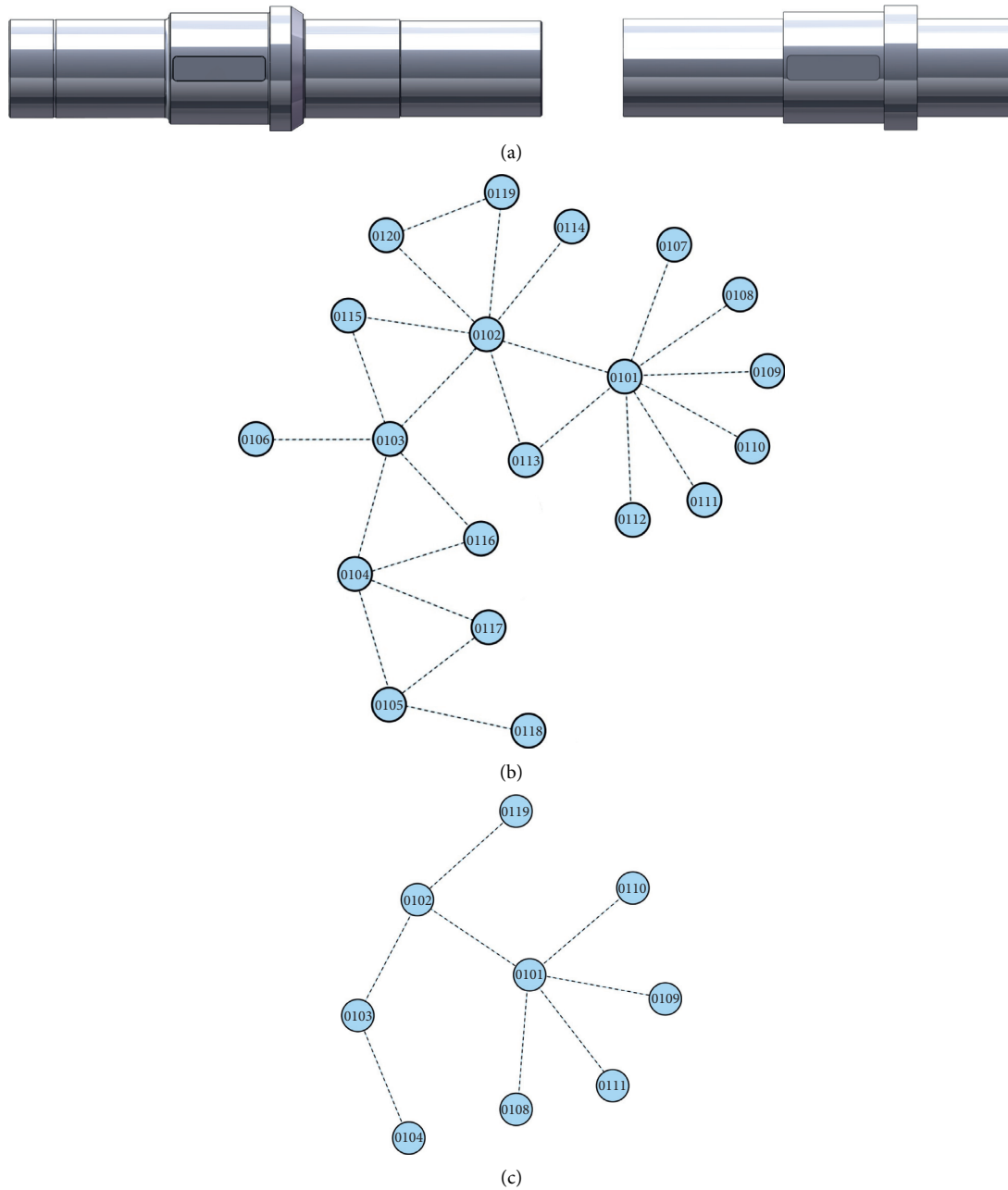


FIGURE 13: Part simplification results. (a) Before and after shaft simplification. (b) Topological relations of features before simplification. (c) Topological relations of features after simplification.

Based on the assembly sequence of geometrically feasible parts, the breadth-first algorithm is applied to obtain three optimal sequences: 1-2-3-4-7-5-9-10-8-6, 1-2-8-6-3-4-7-5-9-10, or 1-8-6-2-3-4-7-5-9-10.

The topological relationship between parts can be represented as the transfer matrix of assembly error

based on the surface by the undirected graph. Taking the axis as an example, the simplified shaft surface relationship is shown in Figure 14(a), while the relationship between the axis surface after adding dimension tolerance and form position tolerance is shown in Figure 14(b). The error transfer model of the assembly

TABLE 7: Assembly sequence planning results.

No. 1	No. 2	No. 3	No. 4	No. 5	No. 6	No. 7	No. 8	No. 9	No. 10	No. 11	No. 12	No. 13	No. 14	No. 15	No. 16	No. 17	No. 18
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	8	8	8
7	7	7	7	7	7	7	7	7	7	8	8	8	8	8	4	4	4
5	5	5	5	5	5	8	8	8	8	6	7	7	7	7	6	7	7
8	8	8	9	9	9	5	5	5	6	7	5	5	5	6	7	5	5
6	9	9	8	8	10	6	9	9	5	5	6	9	9	5	5	6	9
9	6	10	6	10	8	9	6	10	9	9	9	6	10	9	9	9	6
10	10	6	10	6	6	10	10	6	10	10	10	10	6	10	10	10	10
No. 19	No. 20	No. 21	No. 22	No. 23	No. 24	No. 25	No. 26	No. 27	No. 28	No. 29	No. 30	No. 31	No. 32	No. 33	No. 34	No. 35	No. 36
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2	2	8	8	8	8	8	8	8	8
3	3	3	8	8	8	8	8	8	8	2	2	2	2	2	2	2	6
8	8	8	3	3	3	3	3	3	6	3	3	3	3	3	3	6	2
4	4	6	4	4	4	4	4	6	3	4	4	4	4	4	6	3	3
7	7	4	6	7	7	7	7	4	4	6	7	7	7	7	4	4	4
5	6	7	7	5	5	5	6	7	7	7	5	5	5	6	7	7	7
9	5	5	5	6	9	9	5	5	5	5	6	9	9	5	5	5	5
10	9	9	9	9	6	10	9	9	9	9	9	6	10	9	9	9	9
6	10	10	10	10	10	6	10	10	10	10	10	10	6	10	10	10	10

TABLE 8: Tool for the shaft in speed reducer assembly shown in Figure 12.

Part no.	Part name	Tool
1	Shaft	None
2	Key	None
3	Gear	None
4	Locating ring	None
5	End cover	None
6	Sleeves	None
7	Bearing	Tool group
8	Bearing	Tool group
9	Gaskets	None
10	Screw	Screwdriver

TABLE 9: Assembly sequence planning results.

Assembly sequence	Number of assembly direction changes	Number of assembly tool changes	Fitness function value
1-2-3-8-4-6-7-5-9-10	5	5	0.867
1-2-3-4-7-5-9-10-8-6	2	5	0.667
1-8-2-3-4-7-5-9-6-10	4	5	0.867
1-8-2-3-4-7-5-9-10-6	3	6	0.800

product is built after the error transfer model of each part is established. According to the algorithm proposed in the previous paper, the error transfer path in the part is obtained, as shown in Figure 15.

Table 10 shows the results of 10, 100, 1000, and manual dimensional chain calculations. The comparison of data shows that the error obtained by multiple iterations meets the manual calculation results and improves the accuracy.

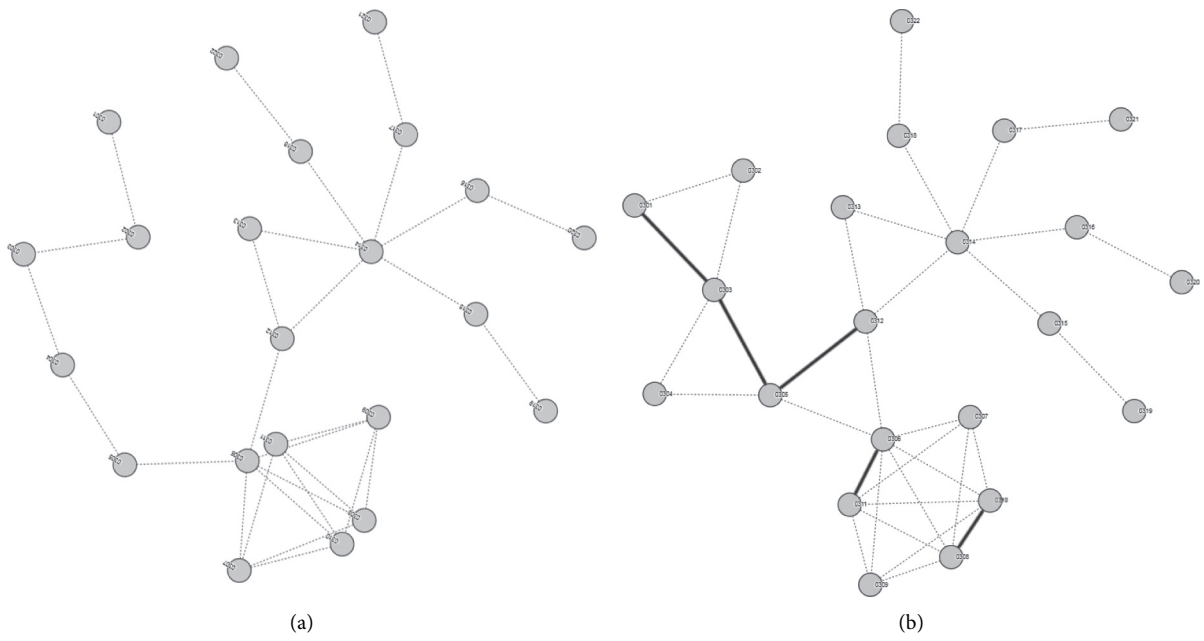


FIGURE 14: Tolerance transfer model of the shaft. (a) Contact relation of the surface in parts. (b) Tolerance relation of the part surface.

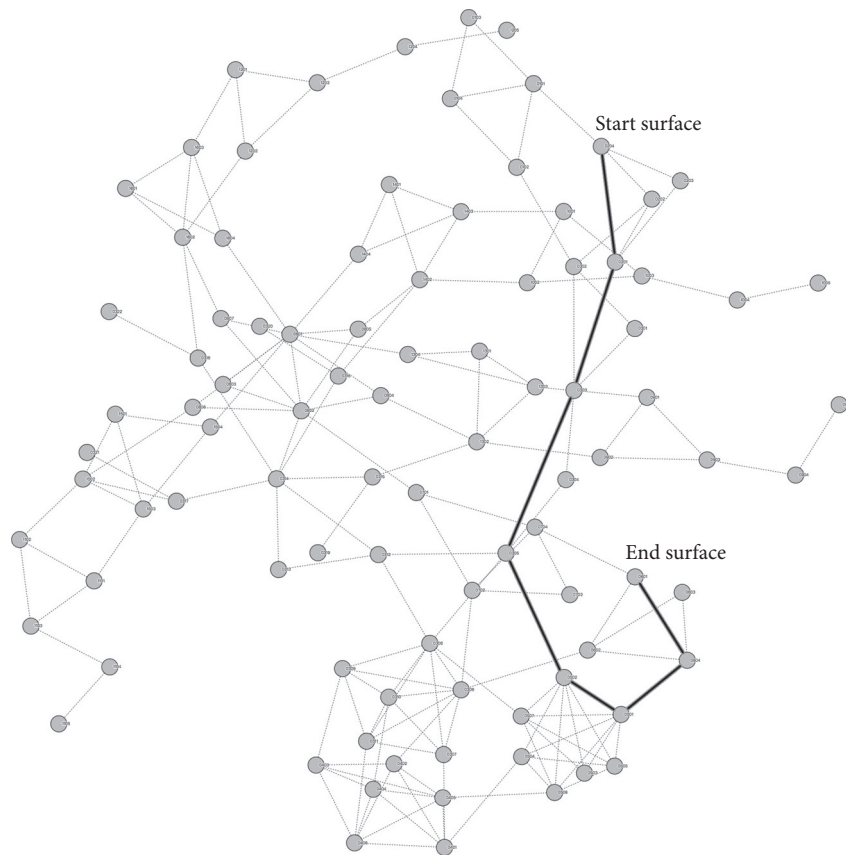


FIGURE 15: Error transmission path.

TABLE 10: Accuracy calculation results.

	10 times	100 times	1000 times	Dimensional chain
d^{sup} (mm)	697.5696	697.4567	697.4403	697.4300
d^{inf} (mm)	697.8537	697.9559	698.0630	698.0660

8. Conclusion

An assembly information model based on the internal structure of parts and the contact relationship among parts is proposed. Firstly, the product is divided into product layer, part layer, feature layer, surface layer, and constraint layer. The assembly information model is built through the topological relationship of features and the contact relationship between parts. On this basis, the assembly constraint model and assembly error transfer model are deduced from the assembly information model. The corresponding relationship between the factors affecting the assembly performance and the total weight is established, and the 3D dimension chain search algorithm is constructed. The main contributions of this study are as follows.

- (1) The assembly information model with multilevel features of the product, part, and assembly process is introduced as a bridge between the product CAD model and the assembly process. High-level semantic concepts of CAD systems and CAD model data of specific products are integrated into a model framework. The assembly information model is built according to the multilevel and multifactor principle, which is the link of CAD and CAPP integration.
- (2) According to the assembly information model, the key features in the assembly process are obtained by using the graph theory to simplify the subsequent calculation complexity.
- (3) According to the contact relationship between parts, the feasible assembly area of parts is deduced, and the assembly information undirected graph is established. By analyzing the influence factors of the assembly information model, the fitness function of the assembly process is constructed.
- (4) The assembly sequence planning and assembly precision calculation are carried out in a model, and the error transfer dimension chain of the assembly is obtained through the sequence planning results. The assembly sequence is taken as the reference factor of precision calculation, which further improves simulation reality. After obtaining the three-dimensional dimension chain of the product, the assembly accuracy of the product can be calculated, including the error transfer and numerical calculation.

Data Availability

The data used to support the findings of the study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

- [1] G. Prabhakaran, P. Asokan, P. Ramesh et al., "Genetic algorithm-based optimal tolerance allocation using a least-cost model," *International Journal of Advanced Manufacturing Technology*, vol. 24, no. 9, pp. 647–660, 2004.
- [2] B. Schleich and S. Wartzack, "A quantitative comparison of tolerance analysis approaches for rigid mechanical assemblies," *Procedia CIRP*, vol. 43, pp. 172–177, 2016.
- [3] Z. Zhu and L. Qiao, "Analysis and control of assembly precision in different assembly sequences," *Procedia CIRP*, vol. 27, pp. 117–123, 2015.
- [4] Y. B. Cai, Z. Q. Cao, and Z. C. Li, "The research of complex assembly dimensional chain analyses and calculation," *Modern Manufacturing Engineering*, vol. 4, pp. 81–84, 2008.
- [5] J. K. Davidson, A. Mujezinović, and J. J. Shah, "A new mathematical model for geometric tolerances as applied to round faces," *Journal of Mechanical Design*, vol. 124, no. 4, pp. 609–622, 2002.
- [6] S. Y. Ding and X. H. Zheng, "Precision control of rotors assembly based on improved Jacobian-Torsor theory," *Acta Aeronautica et Astronautica Sinica*, vol. 42, Article ID 424670, 2021.
- [7] A. Desrochers and A. Clément, "A dimensioning and tolerancing assistance model for CAD/CAM systems," *The International Journal of Advanced Manufacturing Technology*, vol. 9, no. 6, pp. 352–361, 1994.
- [8] T. Jing, X. Tian, X. Liu, H. Hu, M. Zhang, and B. Li, "A multiple alternative processes-based cost-tolerance optimal model for aircraft assembly," *The International Journal of Advanced Manufacturing Technology*, vol. 107, 2020.
- [9] Y. Wang and D. Tian, "A weighted assembly precedence graph for assembly sequence planning," *International Journal of Advanced Manufacturing Technology*, vol. 83, no. 1–4, pp. 99–115, 2016.
- [10] G. Feiyan, L. Jianhua, Z. Fang, Z. Yunong, W. Zhongqi, and L. Shaozhuo, "Research on the state-of-art, connotation and key implementation technology of assembly process planning with digital twin," *Journal of Mechanical Engineering*, vol. 55, no. 17, p. 110, 2019.
- [11] M. Shpitalni, G. Elber, and E. Lenz, "Automatic assembly of three-dimensional structures via connectivity graphs," *CIRP Annals*, vol. 38, no. 1, pp. 25–28, 1989.
- [12] L. S. H. Mello and A. C. Sanderson, "And/or graph representation of assembly plans," *IEEE Transactions on Robotics & Automation*, vol. 6, no. 2, pp. 188–199, 1986.
- [13] M. R. Bahubalendruni and B. B. Biswal, "A review on assembly sequence generation and its automation," *Journal of Mechanical Engineering Science*, vol. 54, pp. 203–210, 2015.
- [14] M. V. A. R. Bahubalendruni, B. B. Biswal, M. Kumar, and R. Nayak, "Influence of assembly predicate consideration on optimal assembly sequence generation," *Assembly Automation*, vol. 35, no. 4, pp. 309–316, 2015.
- [15] L. S. H. de Mello and A. C. Sanderson, "A correct and complete algorithm for the generation of mechanical

- assembly sequences,” *Robotics & Automation IEEE Transactions on*, vol. 7, no. 2, pp. 228–240, 1991.
- [16] U. Thomas, M. Barrenscheen, and F. M. Wahl, “Efficient assembly sequence planning using stereographical projections of C-space obstacles,” in *Proceedings of the IEEE International Symposium on Assembly and Task Planning*, Besancon, France, July 2003.
- [17] J. D. Wolter, “A combinatorial analysis of enumerative data structures for assembly planning,” in *Proceedings. 1991 IEEE International Conference on Robotics and Automation IEEE*, Sacramento, CA, USA, April 2002.
- [18] R. Almgren, “Comparative topological modelling and analysis of assemblies and assembly systems—an aid in computerized assembly planning,” in *Proceedings of the IEEE International Conference on Robotics & Automation*, San Diego, CA, USA, May 2002.
- [19] X. Kou, Y. Cao, Q. Wang, and H. Qiao, “Sub-assembly recognition algorithm and performance analysis in assembly sequence planning,” *International Journal of Advanced Manufacturing Technology*, vol. 107, pp. 1–4, 2020.
- [20] K. Oshihara, T. Hisano, and W. Ueda, “Automatic generation of assembly instructions using step,” in *Proceedings of the 2001 IEEE International Conference on Robotics & Automation*, vol. 3, pp. 21–26, Seoul, Republic of Korea, May 2001.
- [21] C. Pan, S. S. F. Smith, and G. C. Smith, “Determining interference between parts in cad step files for automatic assembly planning,” *Journal of Computing and Information Science in Engineering*, vol. 5, no. 1, pp. 56–62, 2005.
- [22] J. Liu, “The state-of-the-art, connotation and developing trends of the products assembly technology,” *Journal of Mechanical Engineering*, vol. 54, no. 11, p. 2, 2018.
- [23] B. Schleich, N. Anwer, Z. Zhu, L. Qiao, L. Mathieu, and S. Wartzack, “A comparative study on tolerance analysis approaches,” in *Proceedings International Symposium on Robust Design—ISO RD14*, pp. 29–39, Copenhagen, Denmark, August 2014.
- [24] K. Jiang, J. Liu, and R. Ning, “Reasoning and verifying specifications of geometric tolerances based on polychromatic sets theory,” *Computer Integrated Manufacturing Systems*, vol. 21, no. 3, pp. 593–605, 2015.
- [25] A. Clement and P. Bourdet, “A study of optimal-criteria identification based on the small displacement screw model,” *CIRP Annuals-Manufacturing Technology*, vol. 37, no. 1, pp. 503–506, 1988.
- [26] Y. Liu, Z. Wu, and J. Yang, “Mathematical model of size tolerance for plane based on mathematical definition,” *Chinese Journal of Mechanical Engineering*, vol. 37, no. 9, pp. 12–17, 2001.
- [27] A. Clément, A. Rivière, P. Serré, and C. Valade, “The trrss: 13 constraints for dimensioning and tolerancing,” *Geometric Design Tolerancing Theories Standards & Applications*, vol. 59, pp. 122–131, 1998.
- [28] J. K. Davidson and J. J. Shah, “Geometric tolerances: a new application for line geometry and screws,” *Journal of Mechanical Engineering Science*, vol. 216, no. 1, pp. 95–103, 2002.
- [29] H. Xiao and Y. Zhu, “Tolerance conversion of aircraft component intersection axes based on T-map,” *China Mechanical Engineering*, vol. 30, no. 13, pp. 1558–1567, 2019.
- [30] J. Ke, J. K. Davidson, J. J. Shah, and J. Liu, “Using tolerance-maps to transfer datum plane from design tolerancing to machining tolerancing,” in *Proceedings of the ASME International Design Engineering Technical Conferences & Computers & Information in Engineering Conference*, Portland, OR, USA, August 2013.
- [31] W. Liu, L. J. NingRuxin, and K. Jiang, “Precision predicting based on directed deviation graph modeling and D-H methodology,” *Journal of Mechanical Engineering*, vol. 48, no. 7, pp. 125–140, 2012.
- [32] S. Boccaletti, M. Ivanchenko, V. Latora, A. Pluchino, and A. Rapisarda, “Detecting complex network modularity by dynamical clustering,” *Physical Review. E, Statistical, Non-linear, and Soft Matter Physics*, vol. 75, no. 4, Article ID 45102, 2007.
- [33] S. Boccaletti, V. Latora, and Y. Moreno, “Complex networks: structure and dynamics,” *Complex Systems and Complexity Science*, vol. 424, no. 5, pp. 175–308, 2006.
- [34] P. Zhu, J. Yu, and X. Zheng, “Variation propagation network-based modeling and error tracing in mechanical assembling process,” *Journal of Zhejiang University(Engineering Science)*, vol. 53, no. 8, pp. 1582–1593, 2019.
- [35] S. Koo and K. Lee, “Wrap-around operation to make multi-resolution model of part and assembly,” *Computers & Graphics*, vol. 26, no. 5, pp. 687–700, 2002.
- [36] B. C. Kim and D. Mun, “Feature-based simplification of boundary representation models using sequential iterative volume decomposition,” *Computers & Graphics*, vol. 38, pp. 570–577, 2014.
- [37] Y. Kang, B. C. Kim, D. Mun, and S. Han, “Method to simplify ship outfitting and offshore plant equipment three-dimensional (3-D) computer-aided design (CAD) data for construction of an equipment catalog,” *Journal of Marine Science and Technology*, vol. 19, no. 2, pp. 185–196, 2014.
- [38] J. Wang, J. Liu, W. Liu et al., “Modeling technology of tolerance zone in virtual environment,” *Computer Integrated Manufacturing Systems*, vol. 18, no. 12, pp. 231–238, 2012.
- [39] H. Xie and L. Zhong, “Assembly sequence evaluation method based on analytical hierarchy process,” *Computer Systems & Applications*, vol. 21, no. 2, pp. 72–76, 2012.
- [40] Z. Wu, “Stimulated tolerances modeling based on small displacement torsors and tolerances analysis,” *Machinery Design & Manufacture*, vol. 12, pp. 1205–1207, 2010.

Research Article

High Efficiency Spam Filtering: A Manifold Learning-Based Approach

Chao Wang , Qun Li, Tian-yu Ren, Xiao-hu Wang, and Guang-xin Guo

State Grid Beijing Electric Power Company Electric Power Research Institute, Beijing, China

Correspondence should be addressed to Chao Wang; 29681987@qq.com

Received 29 June 2021; Revised 15 July 2021; Accepted 23 July 2021; Published 9 August 2021

Academic Editor: Jie Chen

Copyright © 2021 Chao Wang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Spam filtering, which refers to detecting unsolicited, unwanted, and virus-infested emails, is a significant problem because spam emails lead to unnecessary costs of Internet resources, waste of people's time, and even loss of property. Support vector machine (SVM) is the state-of-the-art method for high accuracy spam filtering. However, SVM incurs high time complexity because of the high dimensionality of the emails. In this study, we propose a manifold learning-based approach for time-efficient spam filtering. From the experiments that most of the features are not decisive, we can obtain the viewpoint that only a minor part of the spam emails can be detected using the nondecisive features. Based on the insight, we propose to employ the Laplace feature map algorithm to obtain the geometrical information from the email text datasets and extract the decisive features. Then, the extracted features are used as the input of SVM to spam filtering. We conduct extensive experiments on three datasets, and the evaluation results indicate the high accuracy time efficiency of our proposed algorithm.

1. Introduction

Email became a popular and widely adopted method in the Internet era since the 1960s for communication, advertisement, and account registration. Spam emails are defined as unsolicited, unwanted, or virus-infested emails [1, 2]. Based on the statistics from Spamlaws, nearly 85% of all emails are spam, in which the advertising, adult-related, and unwanted emails make up 36%, 31.7%, and 26.5% of the content, respectively [3]. Spam filtering, which refers to the process of detecting spam emails, is critical because spam emails are very cheap to send but have severe consequences such as annoying the recipients, wasting the Internet resources, and even leading to loss of property [4]. In spam filtering, nonspam emails should never be classified as spam because the misclassified emails can be critical for the users, which bring significant challenges [5].

The spam filtering methods can be divided into two complementary categories, i.e., origin-based and content-based [6]. In origin-based methods, the senders of the emails are classified as trusted, unknown, and spammer based on the IP addresses, email addresses, allowlists, and blocklists [7]. Emails from trusted senders and spammers will be

directly classified as nonspam and spam, respectively. As for emails from unknown senders, they will be further filtered via content-based methods. This study focuses on the content-based approach in which the classification is purely based on the email content, i.e., header and body.

In the early stage, content-based approaches are mainly based on the statistics of words and phrases in spam and nonspam emails [8]. For example, more than 99% of the emails containing some words and phrases, such as “act now,” “offer expires,” and “winning,” are spam [9]. A spam filter incorporating such statistics is called a Bayesian filter, which classifies the emails by going through the content word by word and phrase by phrase. The advantage of the Bayesian filter is that the classification accuracy can be improved when more data are collected from the users. However, Bayesian filters fail to consider the relationship among the words and phrases, resulting in limited accuracy.

Recently, machine learning becomes popular in content-based spam filtering [10]. Support vector machine (SVM) is one of the successful and cutting-edge techniques achieving higher accuracy than Bayesian filters [11]. SVM embeds the email content into a vector space and separates the emails into two classes, i.e., spam and nonspam, using a hyperplane

in the vector space. The secret of SVM lies in the comprehensive embedding of the email content and separation using a hyperplane. The embedding of the email content incorporates the complex relationship among the words and phrases, which is a comprehensive content representation. The hyperplane separation maximizes the margin between the email embeddings and the hyperplane, making the SVM method robust in spam filtering.

However, embedding of the email content is nontrivial [12–14]. If few features are embedded, the spam filter will incur low classification accuracy; if too many features are embedded, training and applying the spam filter will incur high time overhead. A natural question comes as follows: is it possible to select few features to guarantee high accuracy for SVM-based spam filtering?

In this study, we propose a manifold learning-based approach to select the distinctive features and feed the features to the SVM model for time efficiency and accurate spam filtering. In particular, we gain the insight through experiments that most of the features used in traditional SVM-based spam filtering approaches are not decisive, using which only a minor part of the spam emails can be detected. Based on the insight, we employ an adapted manifold learning algorithm to select the decisive features. Then, the features are fed into the classic SVM model for spam filtering. In this way, our method only selects a small number but decisive features for spam filtering, which provides both high accuracy and time efficiency.

The main contributions of this study are as follows:

- (i) We propose an adapted manifold learning approach to extract the decisive features for spam filtering. The features can not only be used in SVM but also other machine learning-based spam filtering algorithms.
- (ii) We propose a time-efficient SVM-based approach that takes the decisive features as input and filters spam emails
- (iii) We extensively evaluate the proposed spam filtering algorithms, and the experimental results indicate the high accuracy and time efficiency of the proposed method.

The rest of the study is organized as follows. Section 2 presents the related work. Section 3 introduces the proposed method for spam filtering using manifold learning and SVM. Section 4 illustrates the time complexity analysis and demonstrates extensive experimental results. Finally, Section 5 concludes the study with future directions.

2. Related Work

This section presents the related work on machine learning-based spam filtering in Subsection 2.1 and manifold learning in Subsection 2.2.

2.1. Spam Filtering Algorithm. Currently, in the field of spam filtering, the traditional machine learning sorting algorithms include decision-making trees [15], SVM [16], and Bayesian

classifiers. A decision-making tree is a learning algorithm for sorting out datasets based on a tree-like structure. The tree structure includes root nodes and child nodes, representing different attributes of datasets. To form a tree structure is meant to determine the position of different attributes in the decision-making tree, which serves as the learning assignment for the algorithm. Carreras et al. [17] use the decision-making tree model to sort out spam emails, a practice not widely applied to spam filtering for the fact that the attributes of spam emails are hard to be defined. SVM classifiers, however, have pretty wide applications to spam filtering. Its main goal is to learn a linear hyperplane for linear divisible sample point sets and make the sample sets under a given category placed on the one side of the hyperplane while different categories of sample points on the other sides. In training an SVM classifier, only several sample points closest to the linear hyperplane are relevant to model training, and the remaining sample point sets will not work during the training process. Therefore, those several vector-represented sample points that lie closest to the hyperplane are called support vectors. Sculley et al. [18] integrate SVM with the online learning model to filter the spam email. Renuka et al. [19] add latent semantic information in the text message to classification and sort out spam emails using the SVM model.

Undeniably, there are other machine learning classifiers applied to email sorting research, such as ensemble learning algorithm [17], naive Bayesian classifier [20, 21], and reinforcement learning [22, 23]. For small email corpora, some conventional classifier algorithms work effectively in spam filtering. As for massive text corpora, typical machine learning methods are incapable of handling a large amount of text data. Such a backdrop has allowed deep learning techniques to be extensively applied to spam filtering. The deep learning techniques treat an email as a piece of text data from which keywords are extracted for spam identification. Tzortzis et al. [24] initiated a deep learning model for spam filtering, and an autoencoder is employed to detect spams [25].

Although deep learning is emerging and can be applied in spam filtering, deep learning models are not widely accepted by academia and industries for the following two reasons. First, deep learning models are rarely explainable. In spam filtering, nonspam emails should never be classified as spam because the misclassified emails can be critical for the users. Deep learning models can hardly explain the misclassification and are generally not adopted for spam filtering [26]. Second, the training and inference of deep learning models demand a large number of resources (power). The email service providers want to reduce the cost, and deep learning models are not employed [27].

2.2. Manifold Learning. In terms of text classification, such as spam filtering, representation or embedding of the text data is essential to enhance the classification performance. Text representation converts the text data into vector representations that contain necessary information without redundancy and noise. As a result, before training a

classifier, we need to preprocess the text dataset and extract the features for text representation. Such a preprocessing procedure is called feature extraction. Manifold learning is an efficient approach for feature extraction.

Manifold learning is first proposed by Tenenbaum et al. in the Science Magazine in 2000 as a concept of machine learning. Building on the manifold geometric construction, manifold learning is a nonlinear dimensionality reduction technique to reduce high-dimensional samples to nonlinear structure distribution. This kind of algorithm is assumed to place high-dimensional sample point sets on a low-dimensional manifold, which has a highly complicated nonlinear structure and cannot obtain the manifold features as a whole. Therefore, the idea of localization emerges in manifold learning. Essentially, manifold learning is a process of extracting features from high-dimensional datasets, rejecting noise features of no use to a learning assignment, and retaining those useful ones. Therefore, manifold learning is mainly applied to preprocessing datasets, simplifying data representations and reducing time for a learning assignment. At present, there are two types of manifold learning algorithms, i.e., the global structure-preserving dimensionality reduction algorithm, e.g., Isomap [28] containing the geodesic distance between all sample points and local structure-preserving dimensionality reduction algorithms, as shown in locally linear embedding (LLE) [29], Laplacian eigenmaps (LEP) [30], local tangent space alignment (LTSA) [31], and locality preserving projections (LPP) [32].

Text datasets are typically characterized by a highly complex structure in the feature space, and manifold learning can be applied to obtaining the neighboring local structure and the complicated overall structure of datasets. In this way, datasets in the form of text can be processed effectively.

3. Our Proposed Algorithm

This study explores operating the manifold learning algorithm on email text datasets to extract useful features and train the classifier with the SVM algorithm for email classification. This section introduces the manifold learning algorithm and then elaborates on the steps for the mentioned algorithms.

3.1. Laplacian Eigenmaps. The idea of localization is first proposed in manifold learning. That means the critical step to manifold learning is to divide the neighborhood of datasets and then excavate the geometrical characteristics of each neighborhood. In this study, we use the Laplacian eigenmap (LEP) algorithm [30] to extract the useful features of datasets. As a local structure-preserving algorithm, the LEP is time-efficient and features a flexible internal mechanism under which regular terms and other structural information can be added. For the sake of presentation, we suppose the input sample set is expressed as $\{x_1, x_2, \dots, x_N\}$, and after dimensionality reduction, we have the output sample set expressed as $\{y_1, y_2, \dots, y_N\}$. The steps in detail for the LEP algorithm go as follows:

- (i) Step I: calculate the k -neighborhood of all input sample points with the k -nearest neighbor algorithm, and the corresponding neighborhood for x_i is expressed as U_i .
- (ii) Step II: construct the adjacent map on the input sample set and establish the edge structure only between any two points within the sample neighborhood. Each edge is endowed with the weight w_{ij} as

$$w_{ij} = \begin{cases} \exp^{-\left(\|x_i - x_j\|^2 / 2\sigma^2\right)}, & \text{if } x_j \in U_i, \\ 0, & \text{if } x_j \notin U_i, \end{cases} \quad (1)$$

where $\exp^{-\left(\|x_i - x_j\|^2 / 2\sigma^2\right)}$ denotes a Gaussian function, and σ is the parameter of the Gaussian function.

- (iii) Step III: reconstruct a group of sample point sets in the low-dimensional space, so that between the low-dimensional field points, this weight structure can still be satisfied, so the low-dimensional space expression sets can be obtained by optimizing the following function:

$$\min \sum_{ij} \|y_i - y_j\|^2 w_{ij}. \quad (2)$$

Let $Y = [y_1, y_2, \dots, y_N]$ and $[W]_i j = w_{ij}$; the above objective function can be written in the form of matrix representation:

$$\min \text{trace}(YLY^T) \quad \text{s.t. } YDY^T = I, \quad (3)$$

where matrix D is a diagonal matrix, $D_{ii} = \sum_{j=1}^N w_{ij}$, and matrix $L = D - W$ is called the Laplacian matrix.

3.2. Feature Extraction-Based Email Classification Algorithm. Generally speaking, the email datasets we deal with are text datasets. As a result, to fulfill the learning assignment, we need to convert these text datasets into digital ones that algorithms can process. Therefore, the algorithm as a whole has three steps. First, preprocess text datasets; second, use the manifold learning algorithm to extract the features of datasets; finally, operate the SVM algorithm on low-dimensional datasets for classification training. Emails are datasets in the form of text, so we use the doc2vec technique [33] to convert them into datasets in the form of a vector, ensuring machine learning algorithms can process them. Note that doc2vec is a natural language processing tool for representing documents as a vector and is a generalization of the word2vec method [34, 35].

As for a manifold learning algorithm, there are three choices, i.e., Isomap, LLE, and LEP. The Isomap algorithm is inefficient in processing massive ultrahigh-dimensional datasets due to its high time complexity. The LLE algorithm assumes that the local neighborhood of datasets is linear space and acquires the linear correlation representation of the neighborhood. Compared with Isomap and LLE, the LEP algorithm has quite flexible design methods. It works by

constructing the adjacent map of datasets and obtaining the weight on the edge of the map. We can provide additional feature information during this weight acquisition process, like the context structure information of the text and deep layer nonlinear structure information distributed in the text datasets. Therefore, to address the complexity of text datasets, we employ the LEP algorithm to extract text features.

The specific steps for this algorithm are shown in Figure 1 as follows:

- (i) Step I: build an email dictionary. Using the doc2vec technique, we convert the text information of each email into the form of vector representation. After the doc2vec conversion, the initial vector dimensionality can be quite different. To this end, we use the principal component analysis (PCA) [36] to adjust the vector dimensionality and ensure all vector dimensions in good conformity.
- (ii) Step II: reduce, with the LEP algorithm, the dimensionality of the datasets obtained in step I, with specific algorithm steps shown in Subsection 3.1. Moreover, we can obtain the context structure information of the email and apply it to the weight calculation process as regular items. Based on the successive order of sentences and paragraphs in the text, we provide their context weights. Two neighboring sentences boast higher weights than others, and sentences within the same paragraph bear higher weights than those in different paragraphs. Applying this weight information as regular items to the Gaussian function, we can calculate the weight on the adjacent map.
- (iii) Step III: use SVM to classify the datasets of low-dimensional feature representation obtained
- (iv) Step IV: predict the email classification accuracy. We use the method mentioned in Step I to process test emails and convert them into training datasets of the same dimensionality.

4. Analysis and Experiments

The section introduces several types of common datasets in email classification, based on which we compare the accuracy of our new algorithm with conventional classification algorithms to demonstrate the strengths of this new algorithm.

4.1. Time Complexity Analysis. The SVM classifier features the training and testing complexities to be $O(m^2N^2)$ and $O(m^2N)$, respectively, in which N stands for sample point number and m represents the sample feature dimension, following the traditional SVM classifiers [37]. The time complexity of the LEP algorithm involves two parts. On the one hand, k -neighborhood set division takes $O(N^2(m+k))$ time. On the other hand, graph embedding calculation takes $O(dN^2)$ time, where d stands for low-dimensional representation of dimensionality. Therefore, we can use the LEP

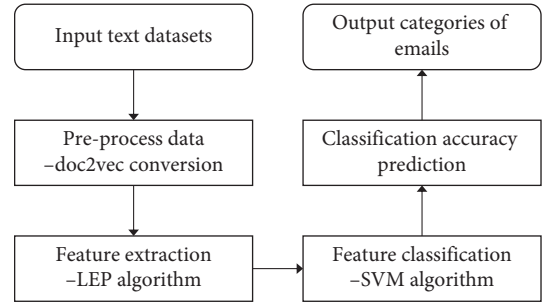


FIGURE 1: Algorithm flow schematic diagram.

algorithm and the SVM algorithm to obtain a new algorithm with time complexity: $O(N^2(m+k+d+d^2))$. If the input sample set dimension m is very high, the new algorithm's time complexity will be approximate $O(N^2m)$. Compared with the SVM algorithm, the time complexity is much lower.

4.2. Spam Datasets. We found there are six representative spam datasets, i.e., EnronSpam, PU1, PU2, PU3, PUA, and GenSpam, as described in Table 1. In particular, EnronSpam [38] is currently a common spam dataset. EnronSpam includes 33,702 emails altogether, including 16,764 regular emails and 16,938 junk emails, accounting for around 50% of the total. With largely the same format comprising subject and text, these emails were mainly from 150 users. These datasets are preprocessed, with each email represented by one text and each text numbered chronologically. This experiment is designed to conduct a contrastive analysis of the algorithm accuracy for these datasets. In terms of PU1, PU2, PU3, and PUA, their content distributions are similar. In this study, we only consider PU1, which is a representative dataset out of the four. To summarize, the performance of our proposed method is compared with the state-of-the-art solutions on three datasets, i.e., EnronSpam, PU1, and GenSpam. We divide each dataset into 70% as the training data and 30% as the test data randomly as usual.

4.3. Experimental Results. This experiment design involves two parts. First, we train classifiers with datasets. Then, we use the trained classifiers to conduct spam email sorting prediction for the test sample sets before calculating their corresponding prediction accuracy. In this experiment, we do the contrastive analysis of the performance of our new algorithm from two aspects, i.e., accuracy prediction and the time taken by the algorithm. Note that the misclassification ratios of the proposed algorithms and the benchmarks are zero through fine calibration. As a result, the misclassification ratio is not included in the comparison results. Specifically, we operate two types of SVM classification algorithms as displayed in [18, 19] on the datasets of PU1 and GenSpam, respectively. Their corresponding accuracy results are given in Table 2. The table provides that the classification accuracy of our SVM + LEP algorithm is not considerably different from that of the other algorithms. With the LEP algorithm applied to the email text's structural

TABLE 1: Several types of spam email corpora.

Corpora	Email	Spam email	Regular email	Spamming rate	Time
EnronSpam [38]	33702	16938	16764	50%	2006
PU1 [39]	1099	484	615	40%	2000
PU2 [39]	721	144	577	20%	2003
PU3 [39]	4139	1821	2318	44%	2003
PUA [39]	1142	570	572	50%	2000
GenSpam [38]	41404	32295	9109	78%	2005

TABLE 2: Different algorithms' classification accuracy.

Datasets	SVM [19]	SVM [18]	SVM + LEP	SVM + LEP + Struc
EnronSpam	92.1%	93.5%	93.9%	94.7%
PU1	95.8%	96.1%	95.6%	96.9%
GenSpam	93.7%	94.6%	93.9%	95.1%

TABLE 3: Time consumption for different algorithms.

Datasets	SVM [19]	SVM [18]	SVM + LEP	SVM + LEP + Struc
EnronSpam	3523 s	2582 s	983 s	1312 s
PU1	743 s	642 s	236 s	285 s
GenSpam	2452 s	1834 s	634 s	715 s

TABLE 4: Different manifold learning algorithms' classification accuracy.

Datasets	Isomap + SVM	LLE + SVM	LPP + SVM	SVM + LEP	SVM + LEP + Struc
EnronSpam	92.4%	93.3%	92.8%	93.9%	94.7%
PU1	94.7%	95.8%	95.1%	95.6%	96.9%
GenSpam	92.9%	93.7%	93.5%	93.9%	95.1%

TABLE 5: Different manifold learning algorithms' time consumption.

Datasets	Isomap + SVM	LLE + SVM	LPP + SVM	SVM + LEP	SVM + LEP + Struc
EnronSpam	6468 s	894 s	965 s	983 s	1312 s
PU1	1683 s	247 s	229 s	236 s	285 s
GenSpam	5478 s	608 s	615 s	634 s	715 s

information, their corresponding classification accuracy improves remarkably.

Now, we analyze the time consumption of different algorithms. We still contrast the two algorithms with our algorithm, whose results are presented in Table 3. The table provides that the time consumption of our new algorithm is significantly lower than that of the other two algorithms. We can use the manifold learning algorithm to extract the features of datasets and effectively reduce the time it takes to train and test these classifiers. Manifold learning also works to remove data noise; so to some extent, it rejects invalid information from the datasets.

4.4. Discussion of Manifold Learning Algorithms. The experiment above helps compare the LEP algorithm-based email classification method with conventional classification methods. Since 2000, manifold learning has given rise to a series of classic algorithms with distinct advantages. In this section, we contrast several types of manifold learning algorithms with the LEP algorithm to show the latter's advantages in email feature extraction. The selected algorithms

are Isomap, LLE, and LPP. In designing the experiment, we have contrastive analysis from two aspects: classification accuracy and time consumption.

We still use three groups of datasets for experimentation on classification accuracy: EnronSpam, PU1, and GenSpam. First, we use different manifold learning algorithms to reduce the dimensionality of the text datasets down to the same low dimension. We then implement the SVM algorithm to fulfill the classification assignment, and the ultimate classification results are presented in Table 4. This table provides that the LEP algorithm with no structural information added sees its dimensionality reduction results similar to those of the other three algorithms. With the structural information added, however, classification accuracy improves significantly. Therefore, in contrast with other manifold learning algorithms, the LEP algorithm has more flexibility in design. It invites additional structural information, while the LPP algorithm learns linear dimension reduction mapping and thus is not as flexible as the LEP algorithm.

In terms of time consumption, we compare different algorithms' time consumption for dimensionality reduction,

and the corresponding experiment contrast results are listed in Table 5. These results show that Isomap has high time complexity, rendering it inefficient in processing massive datasets. Like the LLE and LPP algorithms, the LEP algorithm represents a local dimensionality reduction technique, meaning their time consumption varies barely.

5. Conclusion and Future Directions

Spam filtering has been a critical concern across sectors and industries. If we regard it as a scientific problem, addressing this issue could be a classification issue. However, this classification assignment involves how to have data representation of text and data preprocessing and improve the accuracy of sorting algorithms.

For the top priority of text processing, feature extraction plays a vital role in the follow-up learning assignment. In terms of feature extraction, the typical principal component analysis (PCA) method has some limitations and works only to process the datasets showing the linear structure in data distribution. The standard text datasets themselves have highly complex spatial structures and show high-degree nonlinearity in spatial distribution. With that, it is ineffective to adopt the PCA method for dimensionality reduction. The manifold learning algorithm should be introduced to process the complex structure of text datasets, clean out noise information, and merge redundant features. The objective is to obtain minimum decisive features, reduce data size, and improve learning efficiency.

Manifold learning can obtain the spatial geometric construction of datasets as it bases the geometrical characteristics of datasets on algorithm construction. As for email text datasets, each text comprises a group of sentences, and currently, the mainstream algorithms are learning the feature vector representation of sentences, with few capable of obtaining the spatial structure between sentences. Including manifold assumptions into the algorithm for text analysis, we can increase the structural information of text datasets. Such an advantage is vital to natural language processing, or perhaps to a certain degree, provide some breakthroughs for developing natural language processing-related algorithms. This study is a preliminary attempt to address spam filtering concerns, and in the future, text processing in other fields can also be merged with manifold learning. As long as the complex distributed architecture of datasets is involved, we can always try to bring in the manifold hypothesis to address this problem.

This study introduced some machine learning algorithms in spam sorting. On this basis, we proposed a new learning algorithm, which, together with the manifold learning algorithm, works to preprocess datasets and effectively reduce the algorithm's time complexity. In the Experimentation section, we carried out a contrastive analysis of different algorithms' classification accuracy and time consumption.

In the future, we expect that more machine learning algorithms can be applied to ensuring cybersecurity. Take spam filtering as an example. For the corpora of text, we can introduce the embedding idea in deep learning and use the

embedding method to have a vector representation of the email text. Then, for this type of vector dataset, we can choose some appropriate machine learning algorithms for classification. Unquestionably, the classification results are affected by the embedding performance to add the semantic and structural information of the text to the embedding process. In this way, we may obtain much better classification results.

Data Availability

Some or all data, models, or code generated or used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

- [1] J. Goodman, V. C. Gordon, and D. Heckerman, "Spam and the ongoing battle for the inbox," *Communications of the ACM*, vol. 50, no. 2, pp. 24–33, 2007.
- [2] Y. Jin, T. Li, G. Liang et al., "Spam transaction attack detection model based on gru and wgan-div," *Computer Communications*, vol. 161, pp. 172–182, 2020.
- [3] D. Fetterly, M. Manasse, and M. Najork, "Spam, damn spam, and statistics: using statistical analysis to locate spam web pages," in *Proceedings of the 7th International Workshop on the Web and Databases: Colocated with ACM SIGMOD/PODS 2004*, Paris, France, June 2004.
- [4] C. Godwin and M. Li, "A survey of emerging approaches to spam filtering," *ACM Computing Surveys (CSUR)*, vol. 44, no. 2, pp. 1–27, 2008.
- [5] X. Tian, "A constant time complexity spam detection algorithm for boosting throughput on rule-based filtering systems," *IEEE Access*, vol. 8, pp. 82653–82661, 2020.
- [6] T. S. Guzella and W. M. Caminhas, "A review of machine learning approaches to spam filtering," *Expert Systems with Applications*, vol. 36, no. 7, pp. 10206–10222, 2009.
- [7] O. Amayri and N. Bouguila, "A study of spam filtering using support vector machines," *Artificial Intelligence Review*, vol. 34, no. 1, pp. 73–108, 2010.
- [8] P. P. K. Chan, C. Yang, D. S. Yeung, and W. W. Y. Ng, "Spam filtering for short messages in adversarial environment," *Neurocomputing*, vol. 155, pp. 167–176, 2015.
- [9] H. Shen and Z. Li, "Leveraging social networks for effective spam filtering," *IEEE Transactions on Computers*, vol. 63, no. 11, pp. 2743–2759, 2013.
- [10] T. Gangavarapu, C. D. Jaidhar, and B. Chanduka, "Applicability of machine learning in spam and phishing email filtering: review and approaches," *Artificial Intelligence Review*, vol. 53, pp. 1–63, 2020.
- [11] B. K. Dedetürk and B. Akay, "Spam filtering using a logistic regression model trained by an artificial bee colony algorithm," *Applied Soft Computing*, vol. 91, p. 106229, 2020.
- [12] T. A. Almeida, T. P. Silva, I. Santos, and J. M. Gómez Hidalgo, "Text normalization and semantic indexing to enhance instant messaging and SMS spam filtering," *Knowledge-Based Systems*, vol. 108, pp. 25–32, 2016.

- [13] J. R. Méndez, T. R. Cotos-Yañez, and D. Ruano-Ordás, “A new semantic-based feature selection method for spam filtering,” *Applied Soft Computing*, vol. 76, pp. 89–104, 2019.
- [14] X. Tian and X. Chen, “A weighted feature enhanced hidden Markov model for spam SMS filtering,” *Neurocomputing*, vol. 444, pp. 48–58, 2021.
- [15] S. K. Murthy, “Automatic construction of decision trees from data: a multi-disciplinary survey,” *Data Mining and Knowledge Discovery*, vol. 2, no. 4, pp. 345–389, 1998.
- [16] C. J. C. Burges, “A tutorial on support vector machines for pattern recognition,” *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.
- [17] X. Carreras and L. Marquez, “Boosting trees for anti-spam email filtering,” 2001, <https://arxiv.org/abs/cs/0109015>.
- [18] D. Sculley and G. M. Wachman, “Relaxed online SVMs for spam filtering,” in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 415–422, Amsterdam, the Netherlands, July 2007.
- [19] K. D. Renuka and P. Visalakshi, “Latent semantic indexing based SVM model for email spam classification,” *Journal of Scientific and Industrial Research*, vol. 73, no. 7, 2014.
- [20] V. Metsis, I. Androutsopoulos, and G. Paliouras, “Spam filtering with Naive Bayes-which Naive Bayes?” in *Proceedings of the CEAS Third Conference on Email and Anti-Spam*, vol. 17, pp. 28–69, Mountain View, CA, USA, July 2006.
- [21] I. Androutsopoulos, J. Koutsias, K. V. Chandrinou, P. George, and C. D. Spyropoulos, “An evaluation of Naive Bayesian anti-spam filtering,” 2000, <https://arxiv.org/pdf/cs/0006013>.
- [22] J. Wang, J. Cao, M. Stojmenovic et al., “Pattern-rl: multi-robot cooperative pattern formation via deep reinforcement learning,” in *Proceedings of the 2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 210–215, IEEE, Boca Raton, FL, USA, December 2019.
- [23] Y. Dou, G. Ma, P. S. Yu, and S. Xie, “Robust spammer detection by nash reinforcement learning,” in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 924–933, San Diego, CA, USA, August 2020.
- [24] G. Tzortzis and A. Likas, “Deep belief networks for spam filtering,” in *Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)*, vol. 2, pp. 306–309, Patras, Greece, October 2007.
- [25] G. Mi, Y. Gao, and Y. Tan, “Apply stacked auto-encoder to spam detection,” in *Proceedings of the International Conference in Swarm Intelligence*, pp. 3–15, Springer, Beijing, China, June 2015.
- [26] A. Barushka and P. Hajek, “Spam filtering using integrated distribution-based balancing approach and regularized deep neural networks,” *Applied Intelligence*, vol. 48, no. 10, pp. 3538–3556, 2018.
- [27] S. Srinivasan, V. Ravi, M. Alazab, S. Ketha, A. M. Al-Zoubi, and S. Kotti Padannayil, “Spam emails detection based on distributed word embedding with deep learning,” in *Machine Intelligence and Big Data Analytics for Cybersecurity Applications*, pp. 161–189, Springer, Berlin, Germany, 2021.
- [28] J. B. Tenenbaum, V. De Silva, and J. C. Langford, “A global geometric framework for nonlinear dimensionality reduction,” *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.
- [29] S. T. Roweis and L. K. Saul, “Nonlinear dimensionality reduction by locally linear embedding,” *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.
- [30] M. Belkin and P. Niyogi, “Laplacian eigenmaps and spectral techniques for embedding and clustering,” *Advances in Neural Information Processing Systems (NIPS)*, vol. 14, pp. 585–591, 2001.
- [31] Z. Zhang and H. Zha, “Principal manifolds and nonlinear dimensionality reduction via tangent space alignment,” *SIAM Journal on Scientific Computing*, vol. 26, no. 1, pp. 313–338, 2004.
- [32] X. He and P. Niyogi, “Locality preserving projections,” *Advances in Neural Information Processing Systems (NIPS)*, vol. 16, no. 16, pp. 153–160, 2004.
- [33] Q. Le and T. Mikolov, “Distributed representations of sentences and documents,” in *Proceedings of the 31st International Conference on Machine Learning*, pp. 1188–1196, PMLR, Beijing, China, June 2014.
- [34] T. Mikolov, K. Chen, G. Corrado, and J. Dean, Y. Bengio and Y. LeCun, “Efficient estimation of word representations in vector space,” in *Proceedings of the 1st International Conference on Learning Representations, ICLR 2013, Workshop Track Proceedings*, Scottsdale, AZ, USA, May 2013.
- [35] J. H. Lau and T. Baldwin, “An empirical evaluation of doc2vec with practical insights into document embedding generation,” 2016, <https://arxiv.org/abs/1607.05368>.
- [36] L. I. Smith, *A Tutorial on Principal Components Analysis*, Bibsonomy, Kassel Hessen, Germany, 2002.
- [37] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
- [38] X. Yang, L. Kong, Z. Liu et al., “Machine learning and deep learning methods for cybersecurity,” *IEEE Access*, vol. 6, pp. 35365–35381, 2018.
- [39] A. L. Buczak and E. Guven, “A survey of data mining and machine learning methods for cyber security intrusion detection,” *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1153–1176, 2015.

Research Article

Numerical Investigation of the Trailing Edge Shape on the Added Damping of a Kaplan Turbine Runner

Zhang Ming, P. A. Mbango-Ngoma, Du Xiao-zhen, and Chen Qing-Guang 

College of Mechanical and Electronic Engineering, Shandong University of Science and Technology, Qingdao, Shandong 266590, China

Correspondence should be addressed to Chen Qing-Guang; chenqg@sdust.edu.cn

Received 22 June 2021; Revised 11 July 2021; Accepted 27 July 2021; Published 4 August 2021

Academic Editor: Jie Chen

Copyright © 2021 Zhang Ming et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Hydraulic turbine runners experience high excitation forces in their daily operations, and these excitations may cause resonances to runners, which may induce high vibrations and shorten the runner's lifetimes. Increasing the added damping of runners in water can be helpful to reduce the vibration level during resonances. Some studies have shown that the modification of the trailing edge shape can be helpful to increase added damping of hydrofoils in water. However, the influence of blade trailing edge shape on the added damping of hydraulic turbine runners has been studied in a limited way before. Due to the difficulties to study this problem experimentally, the influence of blade trailing edge shape on a Kaplan turbine runner has been studied numerically in this paper and the one-way FSI method was implemented. The performances of three different turbulence models, including the $k - \epsilon$, $k - \omega$ SST, and transition SST models, in the added damping simulation of the NACA 0009 hydrofoil were evaluated by comparing with the available results of the two-way FSI simulation in the references. Results show that, unlike the significantly different performances in the two-way FSI method, the performances of all the turbulence models are very close in the one-way FSI method. Then, the $k - \epsilon$ turbulence model was applied to the added damping simulation of a Kaplan turbine runner, and results show that the modification of the blade trailing edge shape can be helpful to increase the added damping to some extent.

1. Introduction

Nowadays, hydropower plays an important role in world electricity generation, and over 21% percent of world electricity is produced by hydropower every year [1]. Due to the fast increase of wind and solar energy for electricity generation in recent years, great unstableness is introduced to the electricity grid. Hydropower can provide fast responses for the power regulation to adjust the unstableness of the electricity grid caused by some other renewable sources, like solar and wind energy [1], and its load can change from 20% to 100% in less than 1 minute. Therefore, hydraulic turbines operate at extreme off-design conditions and experience transient events much more times one day than before, which leads to even larger forces. Such higher forces can produce higher vibration levels in the runners, which can cause fatigue damage and shorten their lifetime. Such failures are attributed to the fluid-structure interaction

(FSI) caused by the rotor-stator interaction (RSI) [2–5]. Sometimes, the runner's natural frequency is very much close to the frequency of RSI and/or its harmonic and causes resonances to the runner. This amplifies the amplitudes of runner vibrations and greatly accelerates its fatigue progress [6–8].

Kaplan turbines are one type of widely used hydraulic turbines that are mostly used in low water head and large capacity hydropower plants [9], and the blades of Kaplan turbine runners can rotate to make the runner operate under high efficiencies for a wide range of operation. The typical structure of Kaplan turbines is shown in Figure 1. The excitation forces of Kaplan turbine runners can be both static and dynamic pressures [10–15]. The dynamic pressure mostly comes from the rotor-stator interaction (RSI). Due to the aforementioned reasons, several failure cases for Kaplan turbine runners have been reported in the literature, and some are caused by resonances [6, 16–18].

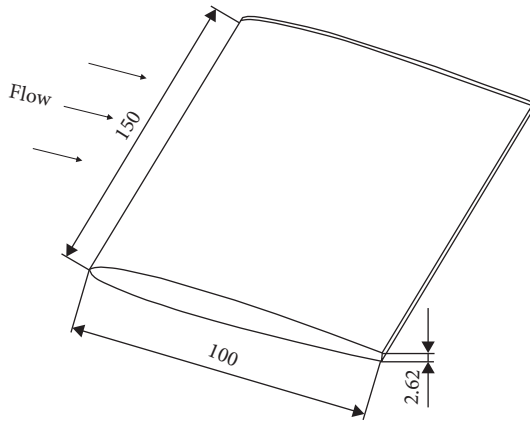


FIGURE 1: Blunt trailing edge hydrofoil.

To reduce the vibration levels of Kaplan turbine runners during resonances to increase their lifetimes, increasing their damping ratios can be helpful. Kaplan turbine runners are submerged in water in daily operations, and their dynamic behavior is severely affected by the added mass and added damping effect from the surrounding water [19–21]. To increase the damping of Kaplan turbine runners, increasing the added damping in water can be a practical way. A good review on the added damping in hydraulic turbines has been shown in [20], where a measure to increase the added damping by modifying the trailing edge shape of the blades has been mentioned. Blake first found that the vibration amplitude of Francis turbine runners can be greatly reduced by sharpening the trailing edge of blades, particularly the Donaldson trailing edge, which has a smooth transition for the sharpening [22]. To study the mechanism of how Donaldson trailing edge reduced the vibration amplitude, Zobeiri investigated the influence of trailing edge shapes, including the blunt trailing edge and the Donaldson trailing edge, on the wake dynamics of a NACA 0009 hydrofoil, and they found that the Donaldson trailing edge can reduce the distance between the vortices after the trailing edge compared with that of blunt trailing edge and increase their collision, which increases the energy dissipation [23]. Yao Yao et al. understand the effect of trailing edge shape from the viewpoint of added damping, and they found that the Donaldson trailing edge can significantly increase the added damping of the NACA 0009 hydrofoil, which explains why the Donaldson trailing edge can be helpful to reduce the vibration level [24, 25]. However, there are no studies on the influence of blade trailing edge shape on the added damping of real hydraulic turbine runners, particularly Kaplan turbine runners.

To measure the added damping of hydraulic turbine runners is very difficult because runners are enclosed by case walls and submerged in water, which makes the sensor installations and data transmissions very difficult. Therefore, investigating the added damping based on numerical methods probably can be an ideal way. The numerical methods for the added damping simulation can be divided into two types, one is the one-way fluid-structure interaction (FSI) method, and another one is the two-way fluid-

structure method. The two-way FSI method needs to give the structure an initial excitation and then measure the vibration amplitude decay ratio to extract the damping ratio. Generally, the two-way FSI method is more theoretical integrity, but the calculation cost is usually very high. Zeng et al. used the two-way FSI method to investigate the influence of trailing edge shape on the added damping of the NACA 0009 hydrofoil, and the results fitted the experimental results very well [25, 26]. In their investigations, they also investigated the influence of turbulence models on the results, including the $k - \epsilon$, $k - \omega$ SST, and transition SST models, and they found that the transition SST model can accurately capture the velocity and pressure distributions in the boundary layer, thus accurate damping compared with experimental results. However, the $k - \epsilon$ model and $k - \omega$ SST model can produce large errors.

The one-way FSI method needs to project the mode shape to the structure boundary and then integrate the flow field parameters under the periodic vibration of the structure boundary. Generally, the one-way FSI method is not theoretical integrity, as the influence of mode shape change due to the fluid flowing is not considered. However, for submerged structures, the mode shape change due to the water flowing usually is not significant, together with considering the much less calculation cost compared with the two-way FSI method, the one-way FSI method is also used by many researchers, and good results were obtained compared with experimental results. Tengs et al. use the one-way fluid-structure method based on the $k - \omega$ SST turbulence model to investigate the added damping of a hydrofoil in flowing water, and little errors were found between the numerical and experimental results [27]. Gauthier et al. developed an improved one-way fluid-structure method with the consideration of the added stiffness and added mass change due to the water flowing and applied this method to investigate the added damping of a hydrofoil and a Kaplan turbine runner [28]. In their study, the $k - \epsilon$ turbulence model was used, and limited errors were found compared with the experimental results.

Due to the too high computation cost of the two-way FSI method, in this study, the one-way FSI method in [28] will be used to investigate the influence of the blade trailing edge shape on the added damping of a Kaplan turbine runner. One problem with the one-way FSI method is the influence of the turbulence model used on the results. Though different turbulence models can produce significant differences in the final results for the two-way FSI method, when different researchers used different turbulence models to do the one-way FSI simulation to simulate the added damping, they all claimed that good results were obtained. Therefore, some uncertainties still lie in the performances of different turbulence models when using the one-way FSI method. To know the performance differences of different turbulence models is important for accurate simulation result achievements. In this study, firstly, the performances of three turbulence models, including the $k - \epsilon$, $k - \omega$ SST, and transition SST models, will be evaluated when they are used in the one-way FSI method for the influence of the trailing edge shape on the added damping of a NACA 0009 hydrofoil

in [25], and then, the turbulence model with the best performance will be applied for the influence of blade trailing edge shape on the added damping of a Kaplan turbine runner.

2. Numerical Methods

The one-way FSI method developed in [28] will be used in this paper. Due to the little influence of damping on the natural frequencies, by neglecting the structure damping, the motion of one vibrating system is described by

$$M_S \ddot{x} + K_S = F_F(t). \quad (1)$$

The natural frequency in vacuum ω_v of the system motion described by equation (1) depending only on structure parameters is

$$\omega_v^2 = \frac{K_S}{M_S}. \quad (2)$$

M_S represents the structure modal mass and K_S the structure modal stiffness, where $F_F(t)$ represents the fluid total modal force applied on the structure and x is representing the deflection.

The structure modal mass and the total force introducing the chosen mode shape are expressed as

$$M_S = \iiint_{\Omega} \rho_S \varphi^2 dV, F_F(t) = \iint_{\Gamma} \tau(t) \cdot \varphi dS. \quad (3)$$

ρ_S represents the structure density, Ω represents the structure volume, and the total surface load induced on the structure by the flow and the fluid-structure interface are, respectively, represented by τ and Γ .

Considering the effect of added mass and added stiffness, the natural frequency, ω_n , of the system becomes

$$\omega_n^2 = \frac{K_S + K_F}{M_S + M_F}. \quad (4)$$

The dimensionless damping ratio is then represented as

$$\zeta = \frac{C_F}{2\omega_n(M_S + M_F)}. \quad (5)$$

M_F represents the fluid-added mass, C_F is the fluid-added modal damping coefficient, and K_F is the fluid-added modal stiffness. The added mass can be interpreted as the mass of fluid accelerated due to the motion of the structure. The added stiffness describes the change in the flow-induced restoring force with the deflection of the structure. The added damping represents energy extracted from the structure as a result of work done by the fluid flow.

Three types of simulations need to be done in order to obtain the above parameters: modal analysis, steady-state CFD analysis, and unsteady CFD analysis. For one selected mode, the modal analysis is mainly used to get the natural frequency, modal mass, and stiffness in vacuum and water. The wet natural frequency and mode shape will act as the initial values for the following simulation. The steady-state CFD is mainly used to get the added stiffness in water, and

the transient CFD is mainly used to update the added mass and calculate the added damping in water.

2.1. Modal Analysis in Vacuum and Water. A modal analysis in vacuum provides the initial circular natural frequency ω_v . The mode shape in vacuum can be extracted, and the modal mass M_S can be calculated from equation (2) or exported from the FEM software directly. The modal stiffness K_S can be calculated from equation (2). Then, the wet modal analysis of the structure in water will be done, which can be achieved through the Acoustic FSI technology available in Ansys. The wet modal analysis will provide the frequency ω_f and mode shape in still water used in the following simulation. The added mass M_F in still water can be calculated from the following equation:

$$\omega_f = \frac{1}{2\pi} \sqrt{\frac{K_S}{M_S + M_F}}. \quad (6)$$

2.2. Steady-State CFD. For the steady state, the modal force can be written as a sum of two forces:

$$F_F = F_0 - K_F x. \quad (7)$$

The fluid-added stiffness, K_F , corresponds to a gradient, defined by the variation of the modal total force over the variation of the deflection $-dF_F/dx$. So, at least two values of both the modal total force and the structure deflection are needed. For linearity purposes, the difference between deflection values has to be very small. URANS flow simulations of five deflection values of the structure were done to get the fluid-added stiffness.

2.3. Transient CFD Analysis. The harmonic modal motion of the structure is defined as

$$x(t) = X_0 \sin(\omega_n t). \quad (8)$$

By projecting the motion to the flow field, a transient CFD simulation can be done. An averaged quantity, Φ , can be calculated as

$$\phi = \frac{1}{N} \int_{t_0}^{t_0 + (2\pi N/\omega_n)} F_F \cdot x(t) dt, \quad (9)$$

where N is an integer number of oscillations and t_0 an arbitrary time coordinate. If the number of periods is high enough, substituting equation (4) into equation (8) results in

$$\phi = \pi X_0^2 \omega_n M_F. \quad (10)$$

From equations (8) and (9), the natural frequency can be updated and used in the following transient CFD simulation. With some repetition, the convergent added mass can be obtained. The average modal work done by the flow on the structure, W , can be obtained as

$$W = \frac{1}{N} \int_{t_0}^{t_0 + (2\pi N / \omega_n)} F_F \cdot \dot{x} dt. \quad (11)$$

If the number of oscillations is high enough, W goes to its convergent value:

$$W = -C_F \pi X_0^2 \omega_n. \quad (12)$$

From equations (11) and (12), C_F can be obtained for the added damping ratio calculation in equation (4).

3. Results and Discussion

3.1. Hydrofoil

3.1.1. Physical Model. The simulated hydrofoil is the 3D NACA 0009 hydrofoil used in [24–26], and in their research, the influence of the trailing edge shape on the added damping was studied. Blunt trailing edge hydrofoil is presented by Figure 1 with a chord length $L = 100$ mm, the width of the span is $w = 150$ mm, and the trailing edge thickness is $h = 3.22$ mm. The hydrofoil of structural steel material has a density of $\rho_S = 7700$ kg/m³, the elasticity modulus is $E = 215$ GPa, and Poisson's ratio is $\nu = 0.3$. The angle of attack of the hydrofoil in the calculation domain is 0° . The configurations of the calculation domain are shown in Figure 2.

The Donaldson trailing edge hydrofoil is a modification of the blunt model. Figure 3 presents the operations to modify the trailing edge from the blunt one to the Donaldson one. Several parameters have to be taken into account as the angle $\alpha = 45^\circ$, the oblique tangent l_1 , and the cubic polynomial curve l_2 . The cubic polynomial is expressed by

$$\begin{cases} y = ax^3 + bx^2 + cx + d, \\ a = 2595.18, b = 637.51, c = 51.141, d = 1.322. \end{cases} \quad (13)$$

3.1.2. Modal Analysis in Vacuum and Water. The modal analysis was done through Ansys Workbench 19.2. Fix support was given to the end of the hydrofoil. For the modal analysis in water, Acoustic FSI technology was used to simulate the added mass effect from the surrounding water, which is a reliable simulation technology used by many researchers [29, 30]. The water domain was selected as the acoustic body with a density of 1000 kg/m³ and a sound speed of 1483 m/s. The inlet and outlet were set to be the total absorption surface, and all other walls were set to be rigid walls. A fluid-structure coupled algorithm was implemented. The mode shape and natural frequency of the first bending mode in water were set as the initial frequency and mode shape for the following CFD analysis. When the hydrofoil is submerged in water, common nodes technology was used for the mesh at the fluid-structure interfaces. The view of the mesh for the hydrofoil in vacuum is shown in Figure 4. The mode shapes of the first bending mode of the blunt and Donaldson hydrofoil in water are shown in Figure 5. The modal parameters are shown in Table 1.

3.1.3. Steady CFD Analysis. The boundary conditions for the CFD analysis are shown in Figure 2. A velocity inlet was used with a low turbulence intensity of 1%, and three different inlet velocities were tested. The outlet boundary is defined as a pressure outlet with a 0 Pa relative pressure. The up, down, left, and right sides are all set as walls. The mode shape profile of the first bending mode was projected to the hydrofoil surface, and in the steady CFD analysis, the hydrofoil was set as the wall with different specific displacements (from 0.01 mm to 0.05 mm). Three different turbulence models, including the $k - \epsilon$, $k - \omega$ SST, and transition SST models, were tested, and for each turbulence model, the modal forces under different hydrofoil displacements were interacted over the hydrofoil surface to calculate the added stiffness. The mesh sensitivity has been strictly checked, and the view of the mesh of the blunt hydrofoil has been shown in Figure 6. The thickness of the first layer of the boundary layers is about 0.001 mm, which corresponds to a y^+ about 1 when the inlet velocity is 15 m/s. For the $k - \epsilon$ turbulence model, scalable wall functions were implemented. A backward Euler second-order time integration scheme was implemented in the analysis. The modal force change versus the hydrofoil displacement under the inlet velocity of 10 m/s has been shown in Figure 7. The modal force nearly linearly increases with the deflection amplitude, and the gradient of the fit curve, that is to say, the added stiffness, is about 0.47. Therefore, the added stiffness is very small compared with the structural stiffness (see in Table 1), and this is maybe because the hydrofoil is symmetric and because of the 0-degree incidence angle of the flow. Due to the small influence of the added stiffness, it is directly neglected for the added damping calculation for all the cases of the hydrofoil.

3.1.4. Transient CFD Analysis. For transient CFD analysis, the mode shape profile of the first bending mode was projected to the hydrofoil surface, and the hydrofoil was set as the wall with periodic displacement with an amplitude of 0.05 mm. The initial frequency is the natural frequency of each hydrofoil in still water. Totally 15 periods of vibration cycles were calculated for each case, and the last 10 periods were used to calculate the added mass and added damping to avoid the influence of the initial turbulence. For each vibration period, 400 substeps were used which has been shown enough for this type of structure [28]. A double-precision solver was selected for calculation via a 64-core cloud computer.

The modal forces along with the time steps (2000 to 4000) for the blunt hydrofoil under the inlet velocity of 5 m/s, 10 m/s, 15 m/s when the turbulence model is transition SST have been shown in Figure 8, and the modal forces along with the time steps (2000 to 4000) for the blunt hydrofoil using three different turbulence models under the inlet velocity of 10 m/s have been shown in Figure 9. When the inlet velocity becomes higher, the modal force curve may become not that smooth which may be because the turbulence intensity becomes higher.

The frequency updating of the Donaldson hydrofoil under the inlet velocity of 10 m/s when using the $k - \omega$ SST

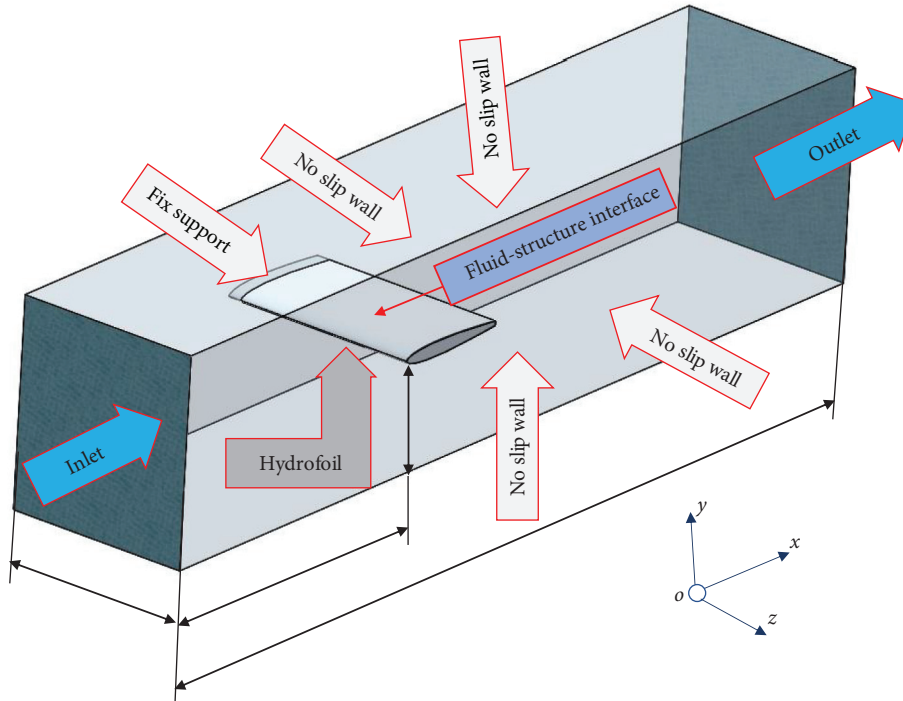


FIGURE 2: Calculation domain configurations of the hydrofoil.

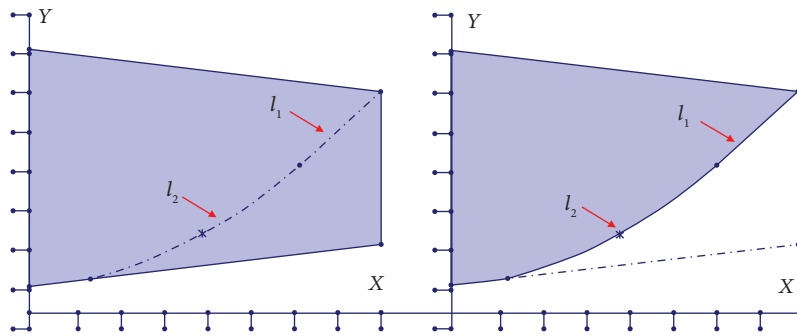


FIGURE 3: Trailing edge modification.

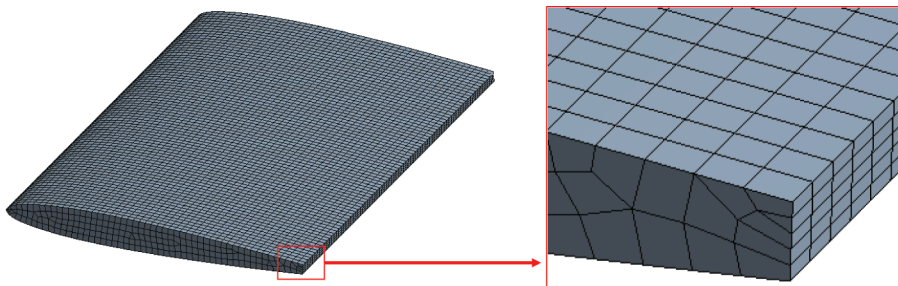


FIGURE 4: Blunt trailing edge hydrofoil mesh.

model has been shown in Table 2. After two steps of updating, the frequency has been nearly convergent. Therefore, two steps of updating were applied to all the simulation cases.

The results obtained from the two-way FSI simulation by [25] are also shown in the figure. As can be seen, the results obtained from different turbulence models are very close for both the blunt and Donaldson hydrofoils, and the results are

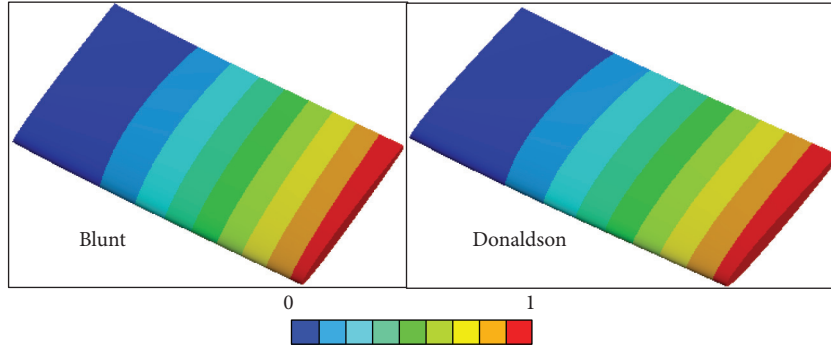


FIGURE 5: Bending mode shapes of the blunt and Donaldson hydrofoil.

TABLE 1: Modal parameter for the bending mode of the blunt and Donaldson hydrofoils.

	f_v (Hz)	f_w (Hz)	M_F/M_S	K_S (N/m)
Blunt	283.26	188.44	2.2596	5076.3
Donaldson	284.38	189.94	2.2416	5157.4

also very close to the results obtained from the two-way FSI simulation. This on the one hand confirms our simulation configurations, and on the other hand, unlike the influence of the turbulence model on the results of the two-way FSI method, the turbulence model seems to affect the results of the one-way FSI method very little. This can also be seen in Figure 7 that the modal forces of the three turbulence models are nearly the same, and thus, nearly the same results are obtained.

3.2. Kaplan Turbine Runner

3.2.1. Physical Model. The investigated turbine is the same one described in [30, 31]. The turbine has a head of 34 m and a maximum power of 73 MW. The runner has 6 blades rotating at 125 rpm, and the distributor has 24 guide vanes. The flow rate at the best efficiency point is about $22.5 \text{ m}^3/\text{s}$. The diameter of the runner is about 6 m, and the tip clearance is 0.09% of the outlet diameter of the runner. The geometry sketch of the studied Kaplan turbine is shown in Figure 11.

Due to the geometry complexity of the blades, it is impossible to sharpen the blade trailing edge with a curve that has an accurate equation as that on the hydrofoil. In this paper, the blade trailing edge was given a proximate 45-degree sharpener with smooth transitions. The geometry view of the blunt and Donaldson blade has been shown in Figure 12.

To purely investigate the influence of the trailing edge shape on the added damping of the runner, only six blades submerged in the water domain were simulated. The blade model has been validated in another paper by the author [30]. For simple, an isolated stage without considering the stay vanes was investigated, and the water domain from the end of the stay vanes to the bottom of the hub was considered. The blade angle of the runner at the nominal load was studied. Only the bending mode of 0 nodal diameters,

i.e., mode for which the motion of every blade is in phase, was considered in the simulation. Due to the low added stiffness of this type of structure [28], the steady CFD analysis was eliminated, and the influence of the added stiffness was neglected.

3.2.2. Modal Analysis. For the modal analysis in vacuum, only one blade was simulated. The blade material is stainless steel. The surface where the rod connects the blade profile was fixed. For the wet modal analysis in water, all the blades were simulated together, and the surrounding water was selected as the acoustic body. The surfaces at which the simulated blades connect with the water domain were set to be fluid-structure interfaces, the inlet and outlet surface were set to be totally absorbed surfaces (the absorption coefficient is 1), and all other surfaces were set to be rigid walls. Only the bending mode was considered. Due to the little mode shape changes from vacuum to water, the blade mode shape in vacuum was used for the following CFD analysis. The view of the mesh in water is shown in Figure 13. The bending mode shapes of the blunt and Donaldson blades have been shown in Figure 14. The natural frequencies of the 0 ND bending modes of the blunt and Donaldson blades in water were selected as their initial frequencies in the following CFD analysis.

3.2.3. Transient CFD Analysis. For the transient CFD analysis, the amplitude of the blade vibration was set to be 0.0005 m, which was also used in [28]. Because the turbulence model has been demonstrated to have little influence on the added damping simulation, the standard $k-\epsilon$ model was used. Scalable wall functions were applied to near-wall regions. Three different inlet velocities, including the 100%, 75%, and 50% of the inlet velocity under the best efficiency of the runner, were simulated. A pressure outlet was also used with a relative pressure of 0 Pa. To simulate the added damping of the 0 ND mode, a Fourier Transformation Method was implemented, which allows us to use only two single blade passages to simulate the entire flow field with different ND blade motions [32, 33]. For every case, a total of 15 periods of oscillation was simulated. The mesh sensitivity was strictly checked, and the view has been shown in Figure 15. For one single blade passage, about 1.2 M

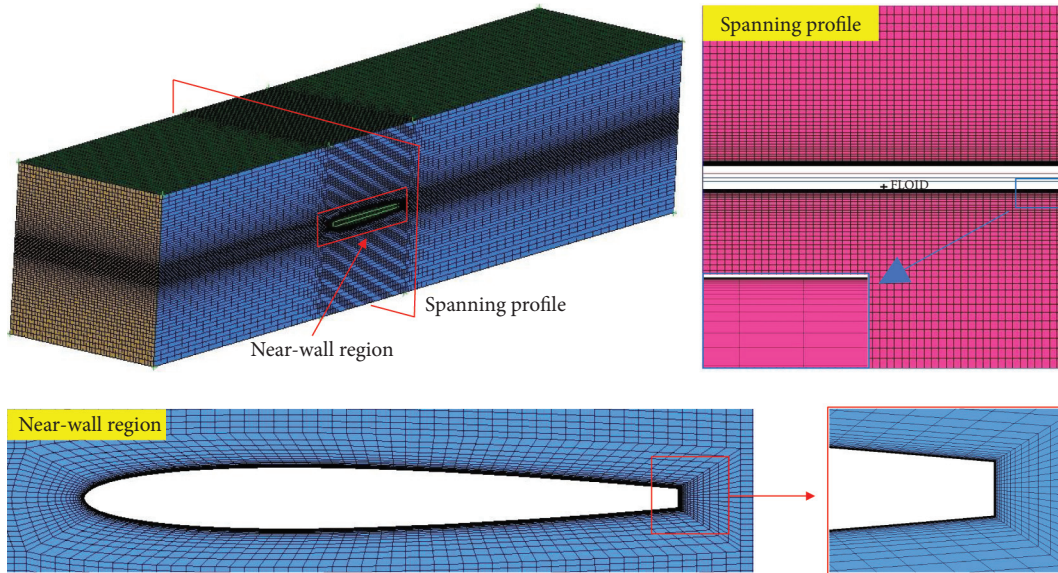


FIGURE 6: Surrounding domain mesh.

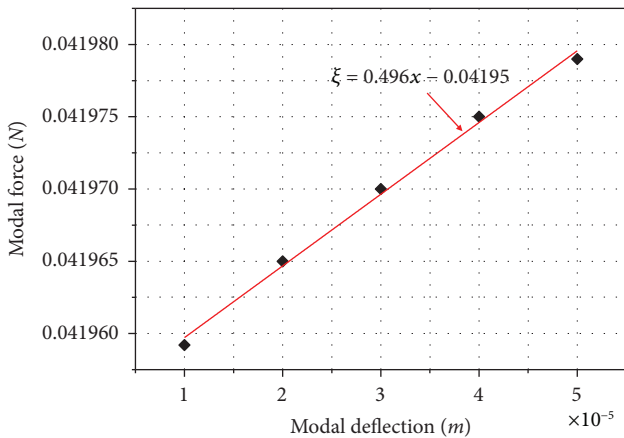


FIGURE 7: Modal force versus the modal deflection in the steady CFD analysis.

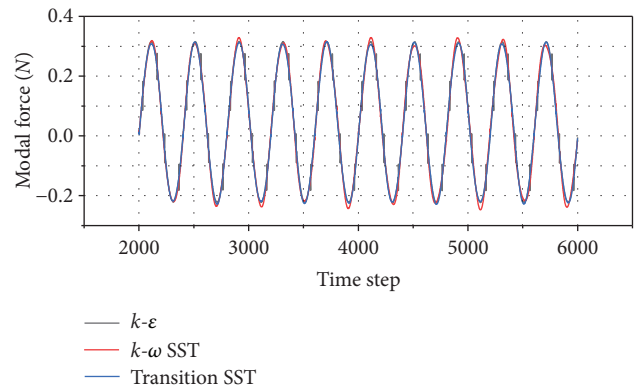


FIGURE 9: Modal forces along with the time steps under different turbulence models.

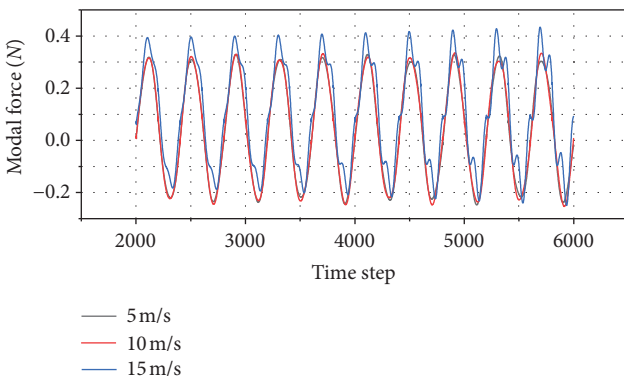


FIGURE 8: Modal forces along with the time steps under different inlet velocities.

unstructured cells were used, and boundary layers on the blade profiles were created with the y^+ lower than 150 (the

TABLE 2: Frequency interactives.

(Hz)	Frequency interactives
Initial frq.	189.94
Step 1	191.26
Step 2	191.41

The final damping ratios of the blunt and Donaldson hydrofoil under different inlet velocities have been shown in Figures 10(a) and 10(b), respectively.

thickness of the first layer is about 0.1 mm). A backward Euler second-order time integration scheme was also implemented in the analysis. The time step was also selected as 400 substeps per period. The double-precision solver was also selected for calculation via the 64-core cloud computer. The same as that on the hydrofoil is also two-step natural frequency updating to get the final added damping. The damping ratios of the blunt and Donaldson Kaplan turbine runners have been shown in Figure 16.

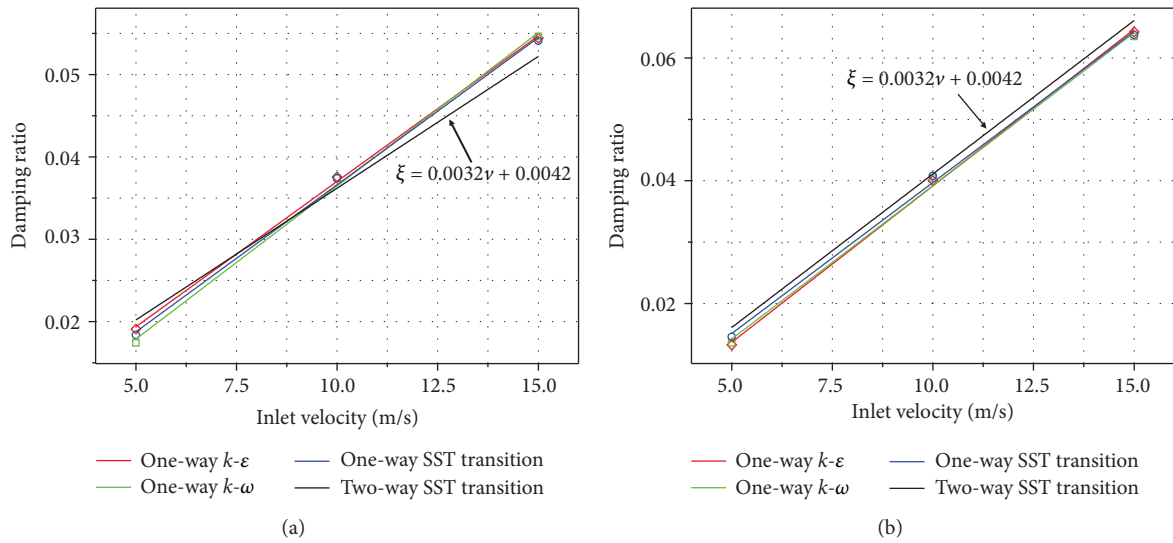


FIGURE 10: Added damping ratio versus the inlet velocity. (a) Blunt hydrofoil. (b) Donaldson hydrofoil compared with the results of the two-way FSI simulation (Zeng, Y. S. et al., 2019).

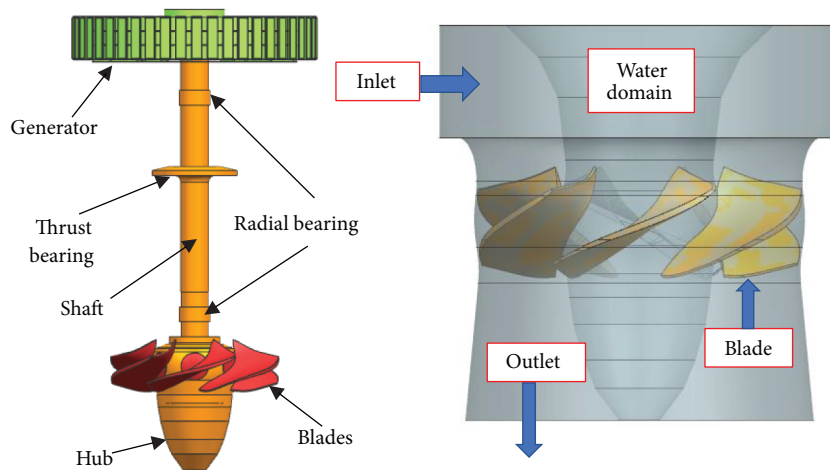


FIGURE 11: Geometry sketch of the studied Kaplan turbine.

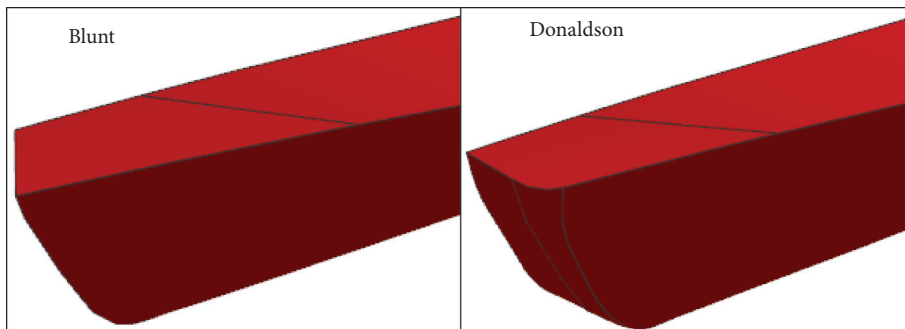


FIGURE 12: View of the blunt and Donaldson trailing edge of the Kaplan turbine blade.

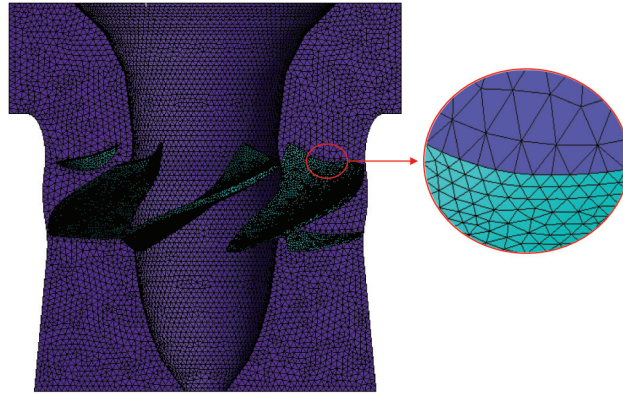


FIGURE 13: View of the mesh when the blades are submerged in water.

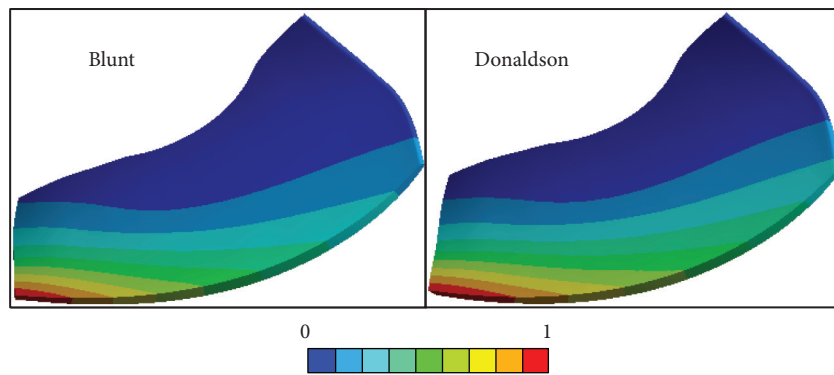


FIGURE 14: Bending mode shapes of the blunt and Donaldson blades.

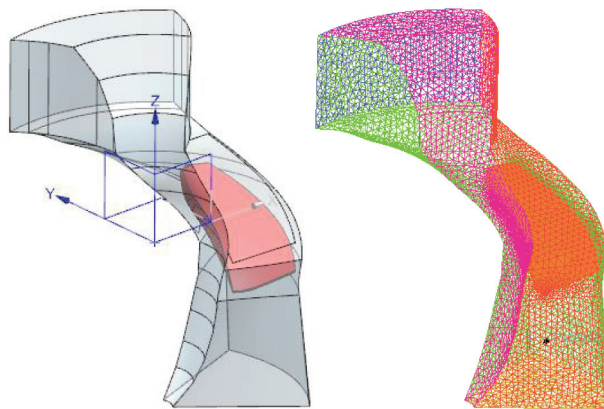


FIGURE 15: View of the mesh of one flow passage.

Same as the hydrofoil, the added damping linearly increases with the inlet velocity for both the blunt and Donaldson trailing edge blades, and generally, the added damping of the Donaldson trailing edge blade is higher than that of the blunt trailing edge blade when the inlet velocity is higher than some values. However, compared with the influence of the trailing edge shape on the added damping, its influence on the Kaplan turbine runner is less significant,

and for each inlet velocity, the difference between two trailing edge shapes is less than 10%. The reason for this may be because the Kaplan turbine blade is wider than the narrow hydrofoil, and the modal displacement concentrates more on other areas of the blade than the hydrofoil, which determines that the trailing edge shape modification produces less significant influence on the modal force than the hydrofoil, thus less significant influence on the added damping.

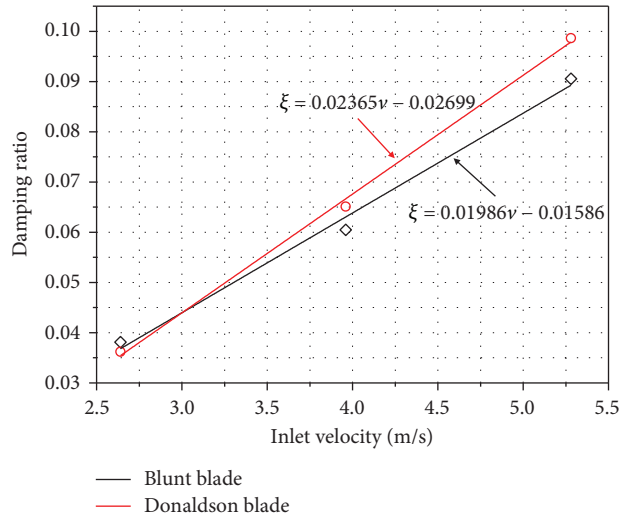


FIGURE 16: The damping ratios of the blunt and Donaldson Kaplan turbine runners.

4. Conclusions

In this paper, the influence of blade trailing edge shape on the added damping on hydraulic turbine runners has been studied. Numerical methods have been used for study in this paper. Because of the too high computation cost of the two-way fluid-structure interaction (FSI) method, the one-way FSI method was implemented. Firstly, the performances of three different turbulence models, including the $k - \epsilon$, $k - \omega$ SST, and transition SST models, in the simulation of the influence of the trailing edge shape on the added damping of hydrofoils, were evaluated by comparing with the results of the two-way FSI simulation available in the references. Unlike the turbulence model which affects the results of the two-way FSI method a lot, the performances of different turbulence models in the on-way FSI method are very close. Little errors found between the results of the three turbulence models are all very close, which provides a reference for the turbulence model chosen in the future added damping simulation using the one-way FSI method. Based on this, the $k - \epsilon$ model was applied to the added damping simulation of a Kaplan turbine runner, and the influence of the blade trailing edge shape on the added damping was investigated. Results show that the Donaldson trailing edge can increase the added damping of the Kaplan turbine runner more than the blunt trailing edge to some extent to reduce the vibration level, but the influence is less significant than that on the hydrofoil. The reason for this may be because the blades of Kaplan turbine runners are wider than the hydrofoil, which determines that the trailing edge shape modification produces less significant influence on the modal force than the hydrofoil, thus less significant influence on the added damping. In the future, the influence of the blade trailing edge shape on the added damping of other types of hydraulic turbine runners which have higher modal displacement concentrations near the blade trailing edge, like the Francis turbine runners, will be investigated to provide suggestions for the vibration level reduction.

Nomenclature

BD: Bending mode
 F_F : Total modal force
 K_F : Added stiffness
 K_S : Structural modal stiffness
 M_S : Structural modal mass
 M_F : Added mass
 n_v : Number of vertical nodal lines
 TS: Torsion mode
 W : Modal work
 Z_v : Number of guide vanes
 φ : Mode shape
 ω_v : Natural frequency in vacuum
 X_0 : Vibration amplitude
 C_F : Added damping
 FSI: Fluid-structure interaction
 k : Flow field pressure pattern
 m_H : Number of horizontal nodal lines
 ND: Nodal diameter
 NL: Nodal line
 RSI: Rotor-stator interaction
 x : Modal deflection
 Z_b : Number of blades
 ω_n : Circular vibration frequency
 ξ : Damping ratio
 ω_f : Natural frequency in vacuum.

Data Availability

Some or all data, models, or codes generated or used during the study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

The authors acknowledge the economic support received from the Natural Science Foundation of Shandong Province (no. ZR2018MEE036) and the Project of Shandong Province Higher Educational Young Innovative Talent Introduction and Cultivation Team (Performance Enhancement of Deep Coal Mining Equipment).

References

- [1] 2018 Iea, Key Electricity Trends 2017 Based on Monthly Data. Monthly Electricity Statistics.
- [2] G. Jacquet-Richardet, M. Torkhani, P. Cartraud et al., "Rotor to stator contacts in turbomachines. Review and application," *Mechanical Systems and Signal Processing*, vol. 40, no. 2, pp. 401–420, 2013.
- [3] C. Nicolet, "Hydroacoustic simulation of rotor-stator interaction in resonance conditions in Francis pump-turbine," *IOP Conference Series: Earth and Environmental Science*, vol. 12, 2010.
- [4] H. D. Li and L. He, "Blade aerodynamic damping variation with rotor-stator gap: a computational study using single-passage approach," *Journal of Turbomachinery*, vol. 127, no. 3, pp. 573–579, 2005.
- [5] G. A. Gerolymos, G. J. Michon, and J. Neubauer, "Analysis and application of chorochronic periodicity in turbomachinery rotor/stator interaction computations," *Journal of Propulsion and Power*, vol. 18, no. 6, pp. 1139–1152, 2002.
- [6] X. Liu, Y. Luo, and Z. Wang, "A review on fatigue damage mechanism in hydro turbines," *Renewable and Sustainable Energy Reviews*, vol. 54, pp. 1–14, 2016.
- [7] S. Vantadori, A. Carpinteri, and D. Scorza, "Simplified analysis of fracture behaviour of a Francis hydraulic turbine runner blade," *Fatigue & Fracture of Engineering Materials & Structures*, vol. 36, no. 7, pp. 679–688, 2013.
- [8] M. Flores, G. Urquiza, and J. M. Rodríguez, "A fatigue analysis of a hydraulic Francis turbine runner," *World Journal of Mechanics*, vol. 2, no. 1, pp. 28–34, 2012.
- [9] H. R. Krishna, *Hydraulic Design of Hydraulic Machinery*, Ashgate Publishing Limited, Farnham, UK, 1997.
- [10] S. Liu, S. Li, and Y. Wu, "Pressure fluctuation prediction of a model Kaplan turbine by unsteady turbulent flow simulation," *Journal of Fluids Engineering*, vol. 131, no. 10, Article ID 101102, 2009.
- [11] L. Zhou, Z. Wang, R. Xiao, and Y. Luo, "Analysis of dynamic stresses in Kaplan turbine blades," *Engineering Computations*, vol. 24, no. 8, pp. 753–762, 2007.
- [12] K. Amiri, B. Mulu, M. Raisee, and M. J. Cervantes, "Unsteady pressure measurements on the runner of a Kaplan turbine during load acceptance and load rejection," *Journal of Hydraulic Research*, vol. 54, no. 1, pp. 56–73, 2015.
- [13] R. Iovănel, D.-M. Bucur, and M. Cervantes, "Study on the accuracy of RANS modelling of the turbulent flow developed in a kaplan turbine operated at BEP. Part 1-velocity field," *Journal of Applied Fluid Mechanics*, vol. 12, no. 5, pp. 1449–1461, 2019.
- [14] R. G. Iovanel, G. Dunca, and M. Cervantes, "Study on the accuracy of RANS modelling of the turbulent flow developed in a kaplan turbine operated at BEP. Part 2-pressure fluctuations," *Journal of Applied Fluid Mechanics*, vol. 12, no. 5, pp. 1463–1473, 2019.
- [15] T. Rus, M. Dular, B. Širok, M. Hočevar, and I. Kern, "An investigation of the relationship between acoustic emission, vibration, noise, and cavitation structures on a kaplan turbine," *Journal of Fluids Engineering*, vol. 129, no. 9, pp. 1112–1122, 2007.
- [16] Y. Luo, Z. Wang, J. Zeng, and J. Lin, "Fatigue of piston rod caused by unsteady, unbalanced, unsynchronized blade torques in a Kaplan turbine," *Engineering Failure Analysis*, vol. 17, no. 1, pp. 192–199, 2010.
- [17] C. V. Frunzaverde and D. Nedelcu, "Failure analysis of a Kaplan turbine runner blade by metallographic and numerical methods," in *Proceedings of the 7th WSEAS International Conference on Fluid Mechanics (Fluids'10)*, Cambridge, UK, January 2010.
- [18] G. Urquiza, J. C. García, J. G. González et al., "Failure analysis of a hydraulic Kaplan turbine shaft," *Engineering Failure Analysis*, vol. 41, pp. 108–117, 2014.
- [19] B. Nennemann, C. Monette, and J. Chamberland-Lauzon, "Hydrodynamic damping and stiffness prediction in Francis turbine runners using CFD," *IOP Conference Series: Earth and Environmental Science*, vol. 49, 2016.
- [20] C. Trivedi, "A review on fluid structure interaction in hydraulic turbines: a focus on hydrodynamic damping," *Engineering Failure Analysis*, vol. 77, pp. 1–22, 2017.
- [21] L. Lu, Y. R. Yang, P. Li, and M. L. Zhang, "Added mass, added stiffness and added damping coefficients for a parallel plate-type structure," *Applied Mechanics and Materials*, vol. 66–68, pp. 1738–1742, 2011.
- [22] W. K. Blake, "Introductory concepts," *Applied Mathematics and Mechanics*, pp. 1–43, 1986.
- [23] A. Zobeiri, "Effect of hydrofoil trailing edge geometry on the wake dynamics," *Ph.D. thesis, EPFL*, Lausanne, Switzerland, 2012.
- [24] Z. Yao, F. Wang, M. Dreyer, and M. Farhat, "Effect of trailing edge shape on hydrodynamic damping for a hydrofoil," *Journal of Fluids and Structures*, vol. 51, pp. 189–198, 2014.
- [25] Y. S. Zeng, Z. F. Yao, P. J. Zhou, F. J. Wang, and Y. P. Hong, "Numerical investigation into the effect of the trailing edge shape on added mass and hydrodynamic damping for a hydrofoil," *Journal of Fluids and Structures*, vol. 88, pp. 167–184, 2019.
- [26] Y. Zeng, "Numerical investigation of added mass and hydrodynamic damping on a blunt trailing edge hydrofoil," *Journal of Fluids Engineering*, vol. 141, no. 8, 2019.
- [27] E. Tengs, "Numerical simulation of the hydrodynamic damping of a vibrating hydrofoil," *IOP Conference Series: Earth and Environmental Science*, vol. 240, 2019.
- [28] J. P. Gauthier, A. M. Giroux, S. Etienne, and F. P. Gosselin, "A numerical method for the determination of flow-induced damping in hydroelectric turbines," *Journal of Fluids and Structures*, vol. 69, pp. 341–354, 2017.
- [29] X. Liu, "Numerical simulation of added mass effects on a hydrofoil in cavitating flow using acoustic fluid-structure interaction," *Journal of Fluids Engineering*, vol. 139, no. 4, 2017.
- [30] M. Zhang, "On the Changes in Dynamic Behavior Produced by the Hydraulic Turbine Runner Damage," 2019.
- [31] M. Zhang, "Influence of internal blade-interactions on the added mass and added damping of a prototype Kaplan turbine runner," *Alexandria Engineering Journal*, 2021, In Press.
- [32] S. Connell, M. Braaten, L. Zori, R. Steed, B. Hutchinson, and G. Cox, "A comparison of advanced numerical techniques to model transient flow in turbomachinery blade rows," *Turbo Expo: Power for Land, Sea and Air*, pp. 1241–1250, 2011.
- [33] S. Gautam, "Numerical investigation of the effects of leakage flow from guide vanes of Francis turbines using alternative clearance gap method," *Journal of Applied Fluid Mechanics*, vol. 13, no. 5, pp. 1407–1419, 2020.

Research Article

Research on Fault Diagnosis of Launch Vehicle's Power Transformation and Transmission System Based on Big Data

Yichi Zhang ¹, Tao Shu,¹ Xincheng Song,¹ Yan Xu,² and Pengxiang Zhang¹

¹Air Defense and Missile Defense College, Air Force Engineering University, Xi'an 710051, China

²Mathematics and Statistics Science College, Shaanxi Normal University, Xi'an 710119, China

Correspondence should be addressed to Yichi Zhang; zyc412181588@ahlcti.com

Received 18 May 2021; Accepted 7 July 2021; Published 22 July 2021

Academic Editor: Jie Chen

Copyright © 2021 Yichi Zhang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The on-board power supply system provides power for the launch vehicle. The power transmission and transformation system plays an irreplaceable role to ensure that the on-board power supply system receives the normal working voltage of the launch vehicle. There are many types of faults in power transmission and transformation systems. The traditional faulty diagnosis method of power transmission and transformation equipment has the disadvantages of being susceptible to experts' subjectivity and model's ossification. In this paper, a new method of equipment fault diagnosis based on big data is proposed. On the basis of big data, this paper introduces the failure mode clustering algorithm, the state parameter correlation analysis algorithm, the fault diagnosis method based on the correlation matrix, and other key fault diagnosis technologies. The fault record data of the 400 kV voltage grade oil-immersed transformer bushing in the past ten years by a Chinese combat unit is used as a case for demonstration. The results show that the accuracy rate of SC-LSTM-K-means clustering model exceeds 95%. And the fault classification mode can be accurately obtained. A priori correlation algorithm with TA coefficient can be used to evaluate the strong and weak relationship between the state parameters; the fault diagnosis matrix based on Pearson's correlation coefficient can accurately determine the fault mode consistent with the actual operation and maintenance test results. Therefore, the fault diagnosis method of power transmission and transformation system based on big data can both effectively obtain the inherent laws of historical data and realize more accurate fault diagnosis with data adaptability.

1. Introduction

The vehicle-mounted power supply system is the power source of the launch vehicle, and the safety of the power transmission and transformation system is the basis of reliable and stable operation of the launch vehicle's power grid, which is of great significance to mobile warfare. Effective and accurate evaluation, diagnosis, and prediction of equipment status can significantly improve the reliability of power supply and the intelligent level of power grid operation [1].

The research on condition monitoring, evaluation, and fault diagnosis technology of high-voltage power equipment was carried out earlier abroad [2, 3]. As early as 1951, engineers from Westinghouse Electric Corporation monitored and diagnosed the motor damage caused by electric discharge in normal operation [4, 5].

Before the 1970s, developed countries such as the former Soviet Union, Japan, the United States, Germany, and Canada made more explorations in live and online monitoring of power transmission and transformation systems. First, they opened up the research field of online monitoring technology and developed partial discharges of dissolved gases in transformer oil, transformers, and gas-insulated switchgear [6]. After the 1990s, equipment condition monitoring and diagnosis technology has developed rapidly, and measurement methods have been continuously improved with the development of sensors, computers, network communication, and other technologies. Monitoring objects have gradually expanded from substation equipment to transmission equipment, and condition information has become increasingly rich. There are also some other test instruments that reflect the equipment condition through

nonelectric quantity measurement, such as ultra-high-frequency partial discharge detection, gas chromatography sensor, optical fiber temperature online measurement, infrared equipment, ultrasonic equipment, etc.

The research on condition monitoring and evaluation of power system equipment in China began in 1970s–1980s [7]. Since the 1980s, the research of online monitoring technology has laid the foundation of development status assessment technology in China. In the past 10 years, live detection and online monitoring systems for primary equipment have been widely used in China. Particularly, with the construction and development of smart grid, online monitoring technology has been rapidly popularized and applied [8, 9]. In recent years, China's power grid companies have made a lot of explorations and attempts in the field of equipment operation and maintenance, gradually realizing the important value of accurately grasping equipment status information, and have begun to promote maintenance management strategies based on status evaluation. In recent years, with the rapid development of sensor technology, countries such as Europe, America, Australia, and Japan have significantly accelerated the research and application of intelligent diagnostic devices. Large foreign companies, especially European companies such as SIEMENS, ABB, Alston, AGE, etc., use online detection systems in high-voltage circuit breakers and GIS. In modern technology, there is still a hard connection between the compartment control cabinet and the primary components of the GIS [10]. In order to overcome this shortcoming, ABB has developed a serial fiber optic bus system, eliminating hard-wired cables and developing the third-generation secondary intelligent technology [11]. ABB adopts intelligent sensor technology and microprocessing technology in the equipment it develops and realizes online monitoring, diagnosis, process monitoring, and in-station computer monitoring of the equipment through digital communication [12].

At present, the widely used state evaluation methods of power transmission and transformation systems in Chinese power grid companies include equipment state scoring system method, expert system method, multidimensional equipment state evaluation method based on traditional machine learning, and sample training method introducing remote expert opinions. However, with the development of intelligent monitoring equipment in recent years, the amount of state parameter data of power transmission and transformation system has increased exponentially; the equipment status data comes from a number of different systems. Traditional state evaluation methods cannot deal with this kind of multisource heterogeneous massive data.

Firstly, this paper analyzes the shortcomings of traditional fault diagnosis methods for power transmission and transformation equipment, including the shortcomings of setting model parameters, being difficult to change after model training and forming, and some relationships being unable to be expressed by equations. A fault diagnosis method of power transmission and transformation system based on big data is proposed. The improved LSTM-K-means algorithm based on silhouette coefficient (SC) is used for fault classification, the a priori correlation algorithm is

combined with TA coefficient to obtain the strong-weak relationship between state parameters, and Pearson's coefficient is used to construct the fault diagnosis matrix. Finally, the feasibility and accuracy of the proposed method are verified by a fault example of 400 kV oil-immersed transformer bushing in combat unit launch vehicle in the recent ten years.

2. State Evaluation Methods of the Traditional Power Transmission and Transformation System

With the development of artificial intelligence algorithms such as neural networks, an equipment state evaluation method based on machine learning has been developed, as shown in Figure 1. This kind of method is usually based on limited sample training and adopts a certain mathematical modeling method to construct a predictable relationship between input and output. Compared with the traditional scoring system, the state evaluation method based on machine learning can use more time section data for sample training, and the prediction results obtained are more accurate than the traditional scoring system. In addition, this method can use complex physical and mathematical functions to model. Usually used mathematical methods include artificial neural network, Bayesian network, support vector machine, and Markov [13].

However, the data that can be used by machine learning method is still limited, and it is still difficult to consider the influence of external factors such as meteorological environment on equipment status in the modeling process of such functions. The model is solidified after setting and training. Unless it is modified or trained again, it cannot adapt to various changes in the process of equipment operation and maintenance, nor can it reflect the influence of differences between different equipment models and different operating environments on diagnosis results. Because the core problem of machine learning method is the selection of historical samples and the training of samples, there are limitations of training speed, training convergence, local minima, and other problems in practical application. Some improvement measures are usually adopted to solve the above problems, such as preimprovement algorithm, postimprovement algorithm, and so on.

On the basis of machine learning method and the concept of expert system, a diagnosis system based on remote expert intervention is developed. The system can train the system with expert opinions as new samples, which integrates the advantages of machine learning method and expert system method, and can improve the accuracy and reliability of subsequent diagnosis results. However, the introduction of remote expert opinions cannot solve the problem of model solidification.

The rise and development of big data mining analysis method has opened up a brand-new technical route for state evaluation and fault diagnosis of power equipment in weapons and equipment and put forward higher requirements for existing equipment state monitoring parameters.

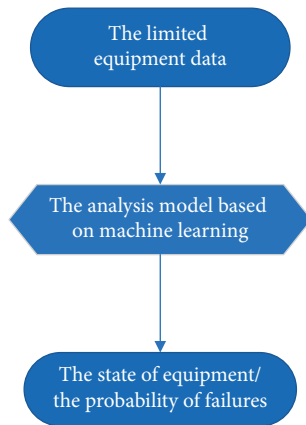


FIGURE 1: Equipment fault diagnosis method based on machine learning.

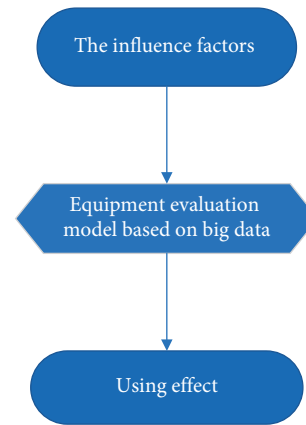


FIGURE 2: Diagram of equipment evaluating model based on big data.

This method introduces the theories and tools of mathematical statistics and pattern recognition. On the basis of large-scale data analysis, it focuses on mining the correlation between analyzed factors under uncertain model conditions. The equipment status evaluation model based on big data is shown in Figure 2. This method adopts the idea of big data mining, focusing on mining and investigating equipment defects and the correlation degree between fault state results and equipment state parameters.

Compared with traditional methods, the most fundamental difference between power equipment condition assessment and fault diagnosis methods based on big data mining analysis is as follows:

- (1) In the evaluation model of traditional methods, the equipment condition monitoring quantity is the input parameter, while the equipment defects and faults are the output parameters; in the analysis method of big data mining, equipment condition monitoring quantity and equipment defects and faults are all input parameters, while output quantity is the association rules, association degrees, and other elements among all input parameters. The traditional model formed by input and output training, once generated, cannot be changed unless retrained; the big data mining analysis method model uses dynamic correlation coefficient matrix to model the correlation between equipment condition index and equipment condition monitoring parameters, which can be continuously regressed and revised and flexibly changed according to the studied equipment objects, state parameters, fault types, etc., without reconstructing the model and without the problem of model solidification. The condition evaluation method of power transmission and transformation equipment based on big data mining is suitable for evaluating and predicting any parameter index of equipment, including equipment health and load capacity.
- (2) In the traditional methods, the evaluation model is the most critical; it is impossible to express the results

other than the preset logical relationship in the evaluation model between the input and output; it is difficult to properly reflect the personalized and differentiated elements such as equipment manufacturers, meteorological environment, habits of operation and maintenance personnel, etc.; however, the correlation mining analysis method based on big data analysis is different from the traditional method. On the basis of higher requirements for existing equipment condition monitoring parameters, it takes massive data as mining objects and uses data mining methods suitable for big data to mine the correlation between factors to be analyzed in uncertain models. In the data mining analysis method, the most important thing is the effective integration and fusion of massive data with multiple sources, multiple time scales, and multiple space-time dimensions, so as to find the inherent (known or hidden) correlation between various equipment condition monitoring quantities and equipment defects and faults, even if it is difficult to integrate the correlation of physical and logical models.

3. Key Technology of System Condition Evaluation Based on Big Data Analysis

3.1. The Clustering Model

3.1.1. K-Means Algorithm. K-means algorithm is proposed by Ding and He [14], which can divide the data into k clusters that minimize the sum of squares of errors through continuous iterative calculation. The algorithm is widely used in all walks of life because of its simple and efficient operation, strong scalability, nearly linear time complexity, and suitability for processing large data sets [15]. The implementation steps of K-means clustering algorithm are as follows [16–18]:

- (1) Initializing the clustering center: randomly selecting K sample points from N sample data as the initial clustering center.

- (2) Cluster division: calculating the distance from the remaining sample points of K initial cluster centers to each initial cluster center and dividing the sample points into the cluster with the smallest distance.
- (3) Calculating a new clustering center: calculating the sum of the divided cluster sample points in step (2), completing the average value calculation, and taking the calculated average value as the new clustering center.
- (4) Convergence judgment: E function is usually used as judgment function, where E function is the sum of squares of errors between sample data and each clustering center. Steps (2) and (3) are circulated until the end of the division cycle that minimizes the E value, which is the best clustering result.

3.1.2. SC-LSTM-K-Means Clustering Model. Based on the long short-term memory (LSTM) network's strong non-linear deep learning capabilities [19] and the advantages of the K -means clustering algorithm, this paper proposes a hybrid clustering model combining LSTM and K -means to identify fault types in power transmission and transformation systems. The flowchart of the LSTM- K -means hybrid model is shown in Figure 3.

In LSTM- K -means clustering, the selection of K value of cluster number is very important. Only by finding out the appropriate K value can we get ideal clustering effect. Silhouette coefficient can solve this problem well. In this paper, by introducing SC, the two concepts of cohesion and separation are integrated, and it is more effective to evaluate the clustering effect by SC.

For the measurement of cohesion in a cluster, the way is to calculate the average distance between the f th element in the cluster and other elements in the cluster, which is denoted as a_i . For the quantification of the separation degree between clusters, the way is to select a cluster b other than the cluster of the above elements, calculate the average value of the distance between the element and all elements in b , then calculate the distance between the above elements and all elements in the other clusters, and find the minimum value of the distance between the element and other clusters, which is recorded as b_i . The equation for calculating the SC_i of the i th element is

$$SC_i = \frac{b_i - a_i}{\max(b_i, a_i)}. \quad (1)$$

Finally, the silhouette coefficients of all elements in all clusters are calculated, and the average value of the silhouette coefficients of each element is obtained as the overall silhouette coefficients of the current cluster.

3.2. Correlation Analysis Algorithm of State Parameters. As far as the current situation is concerned, there are many and complicated equipment state parameters, and there is a lack of mining and analysis of the relationship between equipment parameters, which leads to the lack of systematic

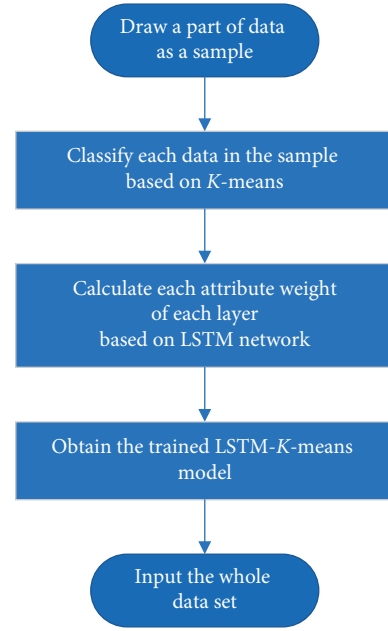


FIGURE 3: The flowchart of LSTM- K -means model.

understanding of equipment parameters. Through the analysis of state parameter association rules, the effective combination of multiple state parameters of equipment, the extraction and merging of feature quantities, and the analysis of the mutual influence degree of state parameters can be realized. For association rules, the general form is the implication of $X \geq Y$, which can be understood as “if X , then Y ,” where X is the equipment state parameter in the preceding item, which can be a single state parameter or a set of multiple state parameters, and Y is the state parameter in the following item, which is generally a single state parameter.

Taking a priori algorithm used in this paper as an example, set the library to be mined of association rules be D , which is a collection of transactions T . If there are n transactions, $D = \{T_1, T_2, \dots, T_j, \dots, T_n\}$, for each transaction, consists of “ m items,” $T = \{I_1, I_2, \dots, I_l, \dots, I_m\}$.

For item set X , the degree of support S is defined as

$$S(X) = \frac{N(X \subseteq T)}{|D|}. \quad (2)$$

For the association rules with $X \geq Y$ $x > y$, the degree of support is

$$S(X \longrightarrow Y) = \frac{N(X \cup Y)}{|D|}, \quad (3)$$

where S is the degree of support and $N(X \subseteq T)$ is the number of sets.

The degree of support described in equation (3) reflects the probability of simultaneous occurrence of these two item sets. The support degree is equal to that of frequent sets.

Similarly, for association rules with $X \geq Y$, its credibility C is

$$C(X \longrightarrow Y) = \frac{S(X \longrightarrow Y)}{S(X)}. \quad (4)$$

The reliability described in equation (4) reflects the probability that if the item set contains X , it also contains Y . For users who use association rules, users can mine association rules with higher S and C by defining thresholds of minimum support and credibility.

3.3. Fault Diagnosis Based on Correlation Matrix. In order to diagnose the fault mode, it is necessary to consider the correlation between each state parameter and each fault mode, that is, the possibility of a certain fault mode when a certain state parameter is abnormal. After obtaining the correlation coefficient between each state parameter and each fault mode of the equipment, the equipment fault mode diagnosis matrix R can be obtained; that is,

$$R = \begin{bmatrix} R_{11} & \cdots & R_{1t} & \cdots & R_{1q} \\ \vdots & & \vdots & & \vdots \\ R_{r1} & \cdots & R_{rt} & \cdots & R_{rq} \\ \vdots & & \vdots & & \vdots \\ R_{p1} & \cdots & R_{pt} & \cdots & R_{pq} \end{bmatrix}, \quad (5)$$

where R_{rt} is the correlation coefficient of vector B_r of the r th failure mode under vector V_t of the t th state parameter; among them, $r \in [1, p]$, there are p fault modes; $t \in [1, q]$, and there are q state parameters. When calculating R_{kt} , there are many sets of data for each failure mode and state parameter, so B_r and V_t are vectors.

When calculating the correlation coefficient R_{rt} , Pearson's coefficient is used in this paper. The correlation coefficient is based on the deviation between the two variables and the average value of their respective variables, calculated by the product-difference method, multiplied by the two deviations, and the product is used to reflect the correlation degree between the two variables [20–22]. Pearson's coefficient ranges from -1 to 1 . A value of 0 indicates that there is no significant linear relationship between the two variables. -1 and 1 indicate that the two variables are completely negative or positive. The following equation shows the correlation coefficient of B_r and V_t applied to equipment fault diagnosis:

$$R_{rt}(B_r, V_t) = \frac{\text{Cov}(B_r, V_t)}{\sqrt{\text{Var}(B_r)} \cdot \sqrt{\text{Var}(V_t)}}, \quad (6)$$

where $\text{Cov}(B_r, V_t)$ is the covariance of B_r and V_t ; $\text{Var}(B_r)$ is the variance of B_r ; and $\text{Var}(V_t)$ is the variance of V_t . After the diagnosis matrix R is obtained by the above method, the fault data can be diagnosed by the following equation:

$$F = R \cdot U. \quad (7)$$

In equation (7),

$$\begin{aligned} U &= [U_1, U_2, \dots, U_t, \dots, U_q]^T, \\ F &= [F_1, F_2, \dots, F_r, \dots, F_p]^T, \end{aligned} \quad (8)$$

where U is the data vector of the fault case to be diagnosed, including the state parameter level of each state parameter,

and F is a fault mode diagnosis result vector, and the value of each element in the vector can indicate the membership degree of the fault case under each fault mode. When the most likely failure mode is finally diagnosed, the failure mode with the largest membership degree (the largest value) can be selected as the final result.

4. Simulation Analysis of Fault Diagnosis Based on Big Data

In this paper, the fault cases of 400 kV oil-immersed transformer bushing in a combat unit launch vehicle in the recent ten years are taken as data mining objects, and the equipment fault diagnosis based on big data mining is studied.

4.1. Preprocessing Data. Firstly, the abnormal state data of the equipment to be mined are collected, with emphasis on the case data of faults and defects. The case code is represented by A_s ($s = 1, 2, \dots, 34$) and the state parameter is represented by W_u ($u = 1, 2, \dots, 26$). According to the representation of faults and defects, the state parameters are assigned. Since the construction of knowledge map is only to mine and analyze the state parameters or equipment abnormal cases themselves, only one state parameter needs to be known whether it is abnormal or not and does not involve the equipment state level or the deterioration degree of the state parameters, so only binary quantification is carried out. According to the severity, defects can be divided into emergency defects, major defects, and general defects.

When a certain state parameter is abnormal, the value of W_u is 1 , which means that the state parameter has faults or major or urgent defects. The value of W_u is 0 , which means that the state parameter is normal. The parameters of the state parameters considered in this paper are shown in Table 1. Through summary statistics, there are 34 groups of equipment failure and defect cases, of which 22 groups are failure data.

4.2. Clustering Analysis of Failure Cases

4.2.1. Clustering Analysis of SC-LSTM-K-Means. Firstly, 22 groups of fault cases in the original data are clustered hierarchically, and all faults can be divided into 3–8 categories by preliminary analysis. K is selected as 3, 4, 5, 6, 7, and 8, respectively, and the clustering effects are compared by calculating the silhouette coefficients of different cluster numbers K . The results are shown in Figure 4. As can be seen from Figure 4, $k = 5$, the clustering result is the most ideal when the failure modes are divided into 5 categories. Combined with the fault physical background and expert experience of transformer bushing, five common fault modes of transformer bushing can be summarized. Using LSTM-K-means to cluster fault cases, the results are shown in Table 2.

4.2.2. Comparison of Three Models' Clustering Accuracy. In recent years, some scholars have studied the combined clustering model of BP neural network and K-means

TABLE 1: Names of status parameters.

State parameter name	Symbol
Porcelain set of creepage	W1
External insulation configuration	W2
Porcelain insulation damage	W3
Oil level indicator	W4
Oil leakage check	W5
Casing grounding	W6
Lead wire at the end of the screen	W7
Infrared temperature measurement	W8
The main insulation dielectric loss and capacitance	W9
Dielectric loss at the end of screen and capacitance	W10
The main insulation resistance	W11
Insulation resistance at the end of screen	W12
Dissolved gas analysis in oil	W13
Outside the insulating oil	W14
Oil dielectric loss	W15
Insulating oil water-soluble acid	W16
Insulation oleic acid value	W17
Insulating oil flash point	W18
Insulating oil moisture	W19
Insulating oil interfacial tension (below 25 degrees Celsius)	W20
Insulating oil breakdown voltage	W21
Insulating oil volume resistance (below 90 degrees Celsius)	W22
Insulating oil sludge and sediment	W23
Insulating oil corrosive sulfur	W24
The carbamate insulating oil	W25
Insulating oil charged tendency	W26

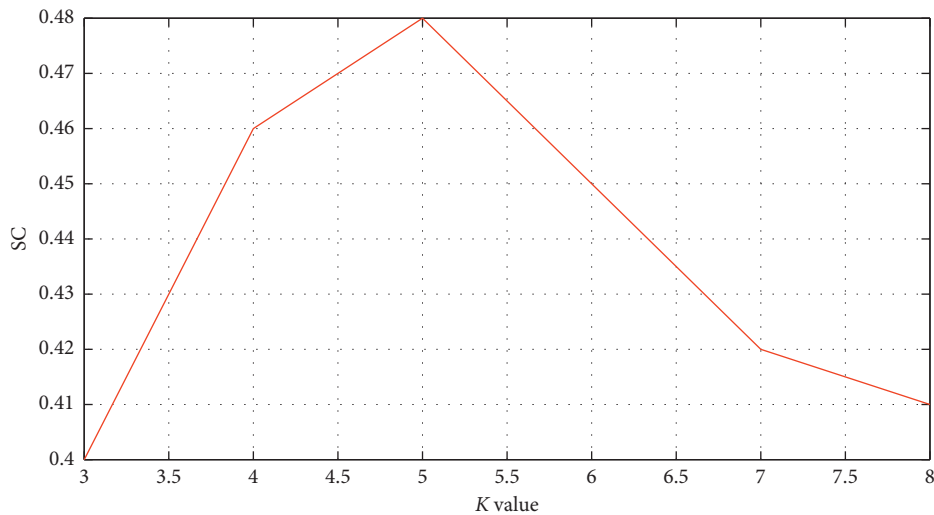


FIGURE 4: Silhouette coefficient under different K values.

TABLE 2: Clustering results of fault data ($k = 5$).

The serial number	Case	Fault name
Failure 1	A4, A6, A11, A12, A19	External insulation affected with damp
Failure 2	A3, A10, A13, A22	Serious oil leakage
Failure 3	A7, A9, A15, A16, A18	Poor contact
Failure 4	A2, A14, A20	Insulating oil affected with damp
Failure 5	A1, A5, A8, A17, A21	Discharge at the end of screen

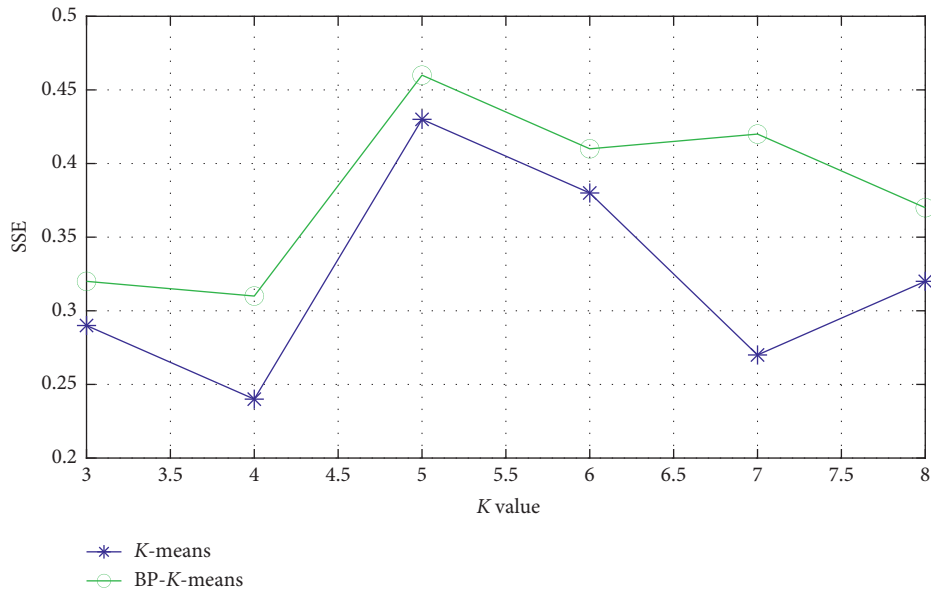


FIGURE 5: K values of K-means and BP-K-means.

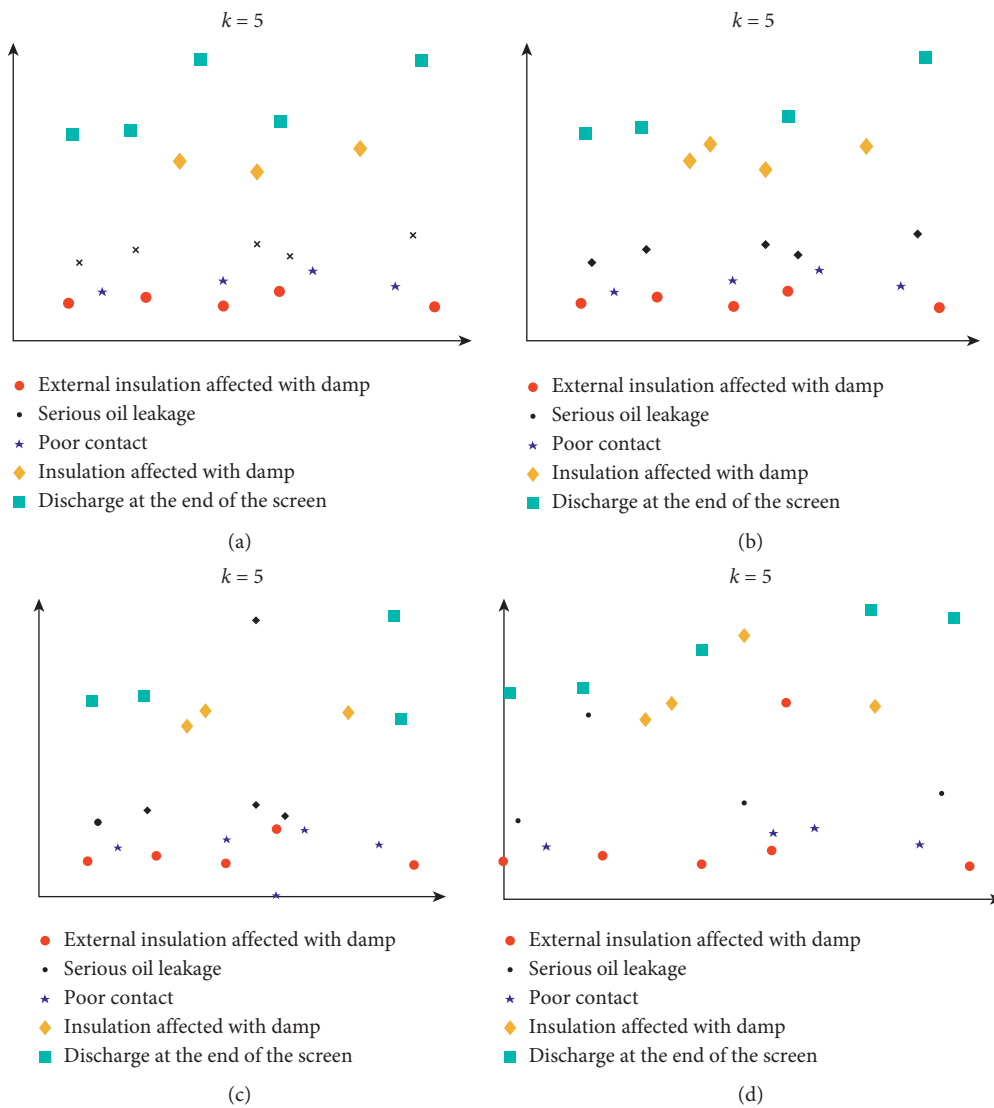


FIGURE 6: Contrast of the three models' clustering results. (a) SC-LSTM-K-means. (b) Real result. (c) K-means. (d) BP-K-means.

TABLE 3: Contrast of the three clustering models.

Models	Indicators		
	Number of correct cases	Number of wrong cases	Accuracy (%)
SC-LSTM-K-means	21	1	95.45
K-means	18	4	81.81
BP-K-means	19	3	86.37

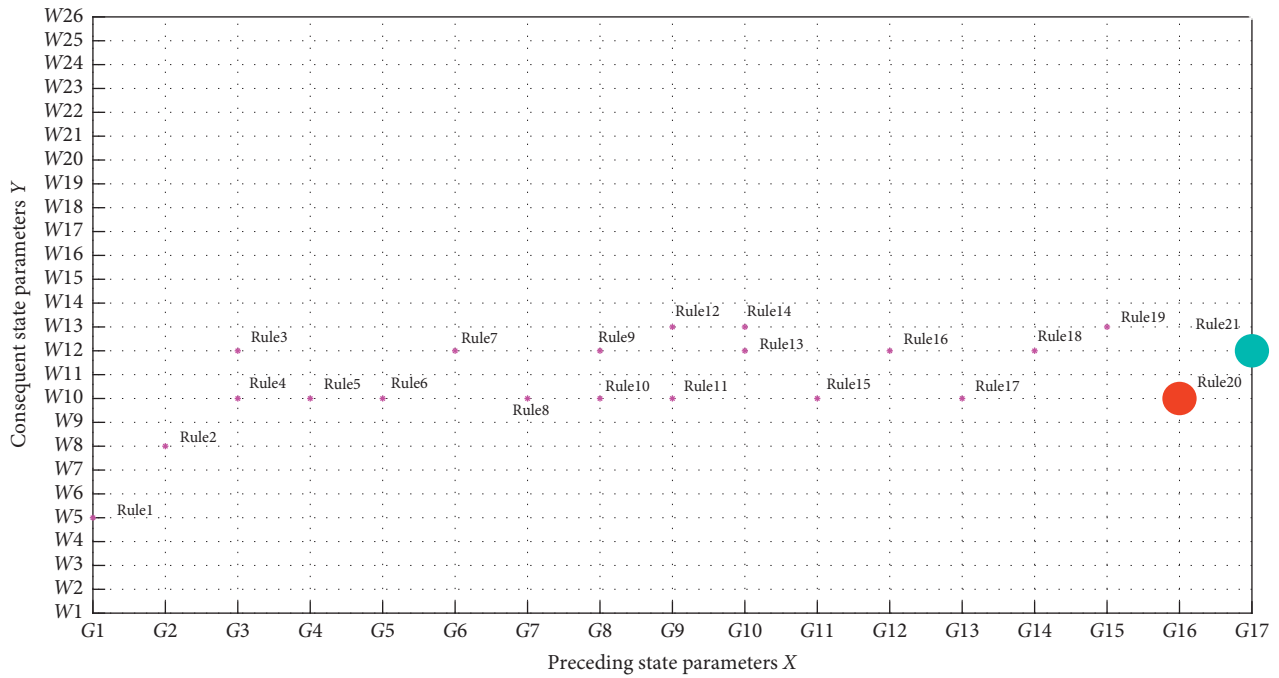


FIGURE 7: Diagram of fault relationship in bushing equipment of oil-immersed transformer.

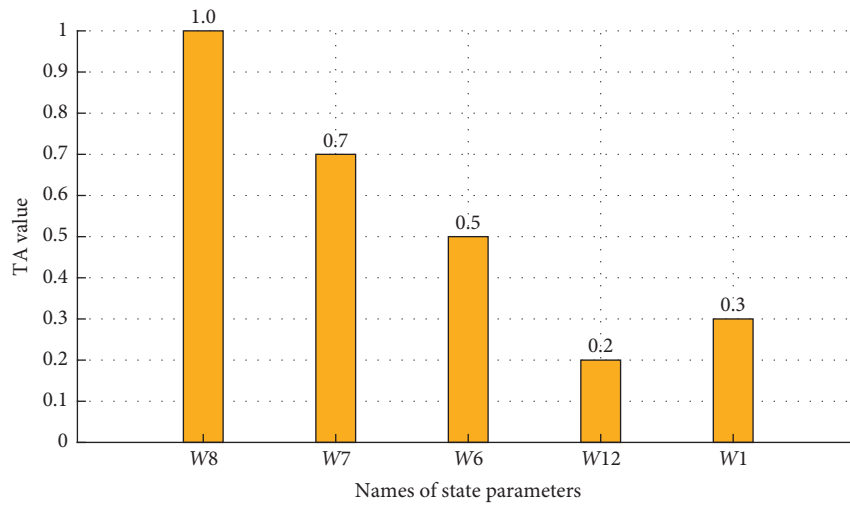


FIGURE 8: TA value of status under poor connection fault.

TABLE 4: Result of fault diagnosis.

Fault name	Fr (case 1)	Fr (case 2)
External insulation affected with damp	-3.1489	-3.0251
Serious oil leakage	2.3013	4.8892
Poor contact	5.1630	-0.7086
Insulating oil affected with damp	-1.3265	-2.2365
Discharge at the end of screen	-0.7208	-0.3216

Calculate the correlation coefficient of each state parameter in the terminal screen discharge fault mode; for example, the vector V_{13} of the state parameter for analyzing the dissolved gas in the oil is

$$V_{13} = [4, 0, 4, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 4, 0, 2, 0, 0, 4, 0, 0]. \quad (10)$$

Therefore, the factor $R_{5\&13}$ in the diagnosis matrix is calculated by

$$R_{5\&13} = \frac{\text{Cov}(S_5, V_{13})}{\sqrt{\text{Var}(S_5)} \cdot \sqrt{\text{Var}(V_{13})}}. \quad (11)$$

Similarly, the remaining elements in the diagnostic matrix R of equation (5) can be calculated. The condition evaluation results of oil-immersed transformer bushing equipment twice are taken as cases for example verification. In case 1, the parameters of casing wiring, terminal screen outgoing line, and infrared temperature measurement state are abnormal; in case 2, the porcelain insulation was damaged, and the parameters of oil level indication and oil leakage inspection state were abnormal. The evaluation of the state parameters is quantified by the deterioration level, and the vectors U of the two cases to be diagnosed are shown in the following equation:

$$U = \begin{cases} [0, 0, 0, 0, 0, 3, 4, 0, 0, 0, 0, 0, 0, 0], & \text{(For case 1),} \\ [0, 0, 2, 3, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0], & \text{(For case 2).} \end{cases} \quad (12)$$

After substituting equation (7), the fault mode diagnosis result vector F of the two diagnosis cases is shown in Table 4.

If the fault mode with the highest diagnosis membership value is taken as the diagnosis result, the diagnosis result of the sample of diagnosis case 1 is poor contact; the diagnosis result of diagnostic case 2 is serious oil leakage. After on-site fault diagnosis by relevant operation and maintenance personnel, the former is an abnormal situation caused by the thread of the casing end screen not tightening, while the latter is an abnormal situation caused by the failure to observe the oil level through the oil level mirror. After inspection, it is caused by the failure to replenish oil for a long time and the normal aging and oil leakage of the casing.

5. Conclusion

Firstly, the SC-LSTM- K -means clustering algorithm can be used to mine the fault modes of power transmission and transformation systems of launch vehicles, and the number of fault classifications can be determined through the SC. By

comparing the clustering effects with K -means and BP- K -means algorithms, it can be concluded that the clustering accuracy of this model reaches 95.45%. Secondly, a priori algorithm based on Boolean association rules to mine frequent item sets can be used to mine the internal correlation relationship of characteristic parameters of power transmission and transformation system, so as to realize the effective combination of multiple state parameters of equipment, feature extraction and merging, and analysis of the mutual influence degree of state parameters; the obtained TA coefficient can characterize the strong-weak relationship of correlation. Finally, the equipment fault correlation matrix based on Pearson's coefficient can get that the maximum value of Fr for case 1 is 5.1630, and the maximum value of Fr for case 2 is 4.8892. The method accurately analyzes and diagnoses the equipment fault mode which is consistent with the actual operation and maintenance test results.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding the publication of this paper.


References

- [1] C. Li, Z. Tang, and Z. Zheng, "Recent progress and future perspective on condition monitoring and diagnostic techniques for power equipment in China," in *Proceedings of the International Conference on Condition Monitoring & Diagnosis*, no. 1, pp. 1-6, IEEE, Perth, Australia, September 2018.
- [2] Z. Liao, F. S. Wen, W. X. Guo et al., "An analytic model and optimization technique based methods for fault diagnosis in power systems," in *Proceedings of the International Conference on Deregulation and Restructuring and Power Technologies*, no. 13, pp. 1388-1393, IEEE, Weihai, China, 2018.
- [3] R. J. Song and X. D. Wang, "Research on assessment and analysis system of condition-based maintenance for power transmission and transformation equipment," *Power System Protection and Control*, vol. 36, no. 9, pp. 54-57, 2018.

- [4] A. Contin, A. Cavallini, G. C. Montanari et al., "Artificial intelligence methodology for separation and classification of partial discharge signals," in *Proceedings of the Conference on Electrical Insulation & Dielectric Phenomena*, vol. 2, no. 09, pp. 522–526, Toronto, Canada, October 2016.
- [5] Y. Jiang and Y. J. Zhou, "Application of distributed partial discharge online monitoring technology in Shanghai 500 kV XLPE power cable line," *High Voltage Engineering*, vol. 41, no. 04, pp. 1249–1256, 2015.
- [6] M. Ge, R. X. Du, G. C. Zhang et al., "Fault diagnosis using support vector machine with an application in sheet metal stamping operations," *Mechanical Systems and Signal Processing*, vol. 18, no. 1, pp. 143–159, 2019.
- [7] J. M. Babita, J. Amit, and B. Srinivasm, "A Web based expert system shell for fault diagnosis and control of power system equipment," in *Proceedings of 2008 International Conference on Condition Monitoring and Diagnosis*, no. 21, pp. 1310–1313, IEEE, Beijing, China, April 2008.
- [8] W. G. Chen, J. Liu, and M. Cao, "Diagnosis method of internal fault for transformers based on information fusion," *High Voltage Engineering*, vol. 41, no. 11, pp. 3797–3803, 2015.
- [9] J. Li, J. Zhang, C. Wang et al., "Research on integrated smart monitoring devices for internet of power transformation equipment," *High Voltage Engineering*, vol. 41, no. 12, pp. 3881–3887, 2015.
- [10] Y. He, "Research on the key technology of power transmission and transformation equipment condition monitoring based on the Internet of Things," Thesis, Hunan University, Hunan, China, 2014.
- [11] L. Qin, "Research and application of fault diagnosis method for power transmission and transformation equipment based on online monitoring," Thesis, North China Electric Power University, Beijing, China, 2014.
- [12] M. M. Lu, "Research on fault rate model and fault diagnosis of oil-immersed transformer," Thesis, Zhejiang University, Zhejiang, China, 2014.
- [13] Z. L. Wu, "Research oil fault diagnosing methods for power transformers," Thesis, North China Electric Power University, Beijing, China, 2018.
- [14] C. Ding and X. He, "K-means clustering via principal component analysis," in *Proceedings of the Twenty-First International Conference on Machine Learning ACM*, no. 3, pp. 15–21, Da Lat, Vietnam, January 2019.
- [15] L. Han, Y. Qiao, and H. T. Jing, "Wind power interval assessment based on error classification," *Automation of Electric Power Systems*, vol. 45, no. 1, pp. 97–104, 2021.
- [16] L. H. Zhang, "Research and application of K-means clustering algorithm," Thesis, Wuhan University of Technology, Wuhan China, 2017.
- [17] S. K. Sun, C. Dong, Z. Wang et al., "Wind power medium and long-term electricity forecasting methods considering different wind energy characteristics," *High Voltage Technology*, vol. 47, no. 4, pp. 1224–1233, 2021.
- [18] G. Q. Weng, Y. G. Gong, J. P. Shu et al., "Active distribution network power quality prediction based on clustering LSTM deep learning model," *High Technology News*, vol. 30, no. 7, pp. 687–697, 2020.
- [19] R. Sibson, "SLINK: an optimally efficient algorithm for the single link cluster method," *Computer Journal*, vol. 16, no. 1, pp. 30–34, 2010.
- [20] X. M. Li and W. H. Jiang, "Research on linear correlation and dependence measure," *Mathematics in Practice & Theory*, vol. 36, no. 2, pp. 188–192, 2016.
- [21] M. W. Lin, C. Huang, R. Q. Chen et al., "Directional correlation coefficient measures for Pythagorean fuzzy sets: their applications to medical diagnosis and cluster analysis," *Complex & Intelligent Systems*, preprint, 2021.
- [22] I. H. Hariyanto, D. G. Pratomo, and M. R. Darminto, "The correlation of tsunami heights and coastal conditions in Palu Bay using the contingency coefficient analysis," *IOP Conference Series: Materials Science and Engineering*, vol. 1052, no. 1, 2021.

Research Article

Fire Prediction Based on CatBoost Algorithm

Fangrong Zhou,¹ Hao Pan,¹ Zhenyu Gao,¹ Xuyong Huang,¹ Guochao Qian,¹ Yu Zhu,² and Feng Xiao³ 

¹Joint Laboratory of Power Remote Sensing Technology, Electric Power Research Institute, Yunnan Power Grid Company Ltd., Kunming 650217, China

²China Academy of Space Technology Institute of Spacecraft System Engineering, Beijing 100094, China

³School of Computer Science, Guangdong University of Technology, Guangzhou, China

Correspondence should be addressed to Feng Xiao; 2112005051@mail2.gdut.edu.cn

Received 17 May 2021; Revised 22 June 2021; Accepted 12 July 2021; Published 20 July 2021

Academic Editor: Jie Chen

Copyright © 2021 Fangrong Zhou et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In recent years, increasingly severe wildfires have posed a significant threat to the safe and stable operation of transmission lines. Wildfire risk assessment and early warning have become an important research topic in power grid risk assessment. This study proposes a fire prediction model on the basis of the CatBoost algorithm to effectively predict the fire point. Five wildfire risk factors, including vegetation factors, meteorological factors, human factors, terrain factors, and land surface temperature, were combined using the feature selection method on the basis of the gradient boosting decision tree model and principal component analysis to achieve dimensionality reduction of redundant data and create a fire prediction model. The MODIS fire point product is used as the model evaluation data. The verification result uses the AUC value as the evaluation factor. The accuracy of the model is 0.82, and the AUC value is 0.83. The obtained fire point evaluation results are in good agreement with the actual fire points. Results show that this model can effectively predict the risk of wildfires.

1. Introduction

Mountain fire disaster is an essential factor that destroys the forest ecosystem and affects the safe and stable operation of the power grid [1, 2]. Mountain fires accounted for 60% of all the emergencies that have changed the stable operation of the power grid in recent years [2]. According to statistical analysis over the years, most reclosing of transmission line trips caused by wildfire disasters will fail, which seriously affects the quality of life in the area and causes substantial economic losses to relevant departments.

Most regions in southern China are located in forest areas, with dense forests, complex terrain, and dry climate, which provide a good material basis for the occurrence of mountain fire disasters, leading to frequent mountain fire disasters and posing a considerable threat to the safe and stable operation of the power grid [3, 4]. Mountain fire disasters have become an important factor that affects the safe and stable operation of the power supply system.

Therefore, effectively predicting the fire risk of woodland, grassland, and cultivated land that may occur in the future and making corresponding warnings are considerably significant to maintain the stable operation of the power grid [5].

Scholars at home and abroad are mainly divided into two directions in the research of wildfire risk: purely using meteorological data for wildfire risk assessment and combining the tripping mechanism of transmission lines, vegetation factors, and human factors to classify wildfire risk levels. At present, meteorological departments and forestry departments mainly assess the risk of wildfires from the perspective of meteorology [6]. In 1995, Wang et al. [7] and others proposed a new technology for forest fire risk assessment based on meteorological elements such as temperature, humidity, precipitation, and wind speed, but it is only suitable for large-area forest fire risk forecasting. Literature [8] built a graph model-based overhead transmission line wildfire risk prediction model based on the

meteorological factors, combined with surface combustion factors and historical fire factors. This method has been effectively applied to a certain southern power grid. Literature [9] uses forest fire danger meteorological grades to assess the probability of wildfires and establishes a risk assessment model for transmission lines with temporal and spatial distribution characteristics. Literature [10] established a risk assessment model from two aspects: the risk of wildfire disasters and the vulnerability of transmission lines. Literature [11] combined the relationship between normalized differential vegetation index (NDVI), satellite remote sensing fire point, rainfall, and other factors with the occurrence of wildfires on transmission lines and proposed a wildfire risk assessment model for transmission lines, but only monthly risk assessment. In fact, fires are very closely related to human activities. Literature [11] proposed a fire prediction model that combines meteorological data and human activities. The model is applied in areas with severe fire disasters, and it has good prediction accuracy. Literature [12], based on historical meteorological data, vegetation, data and terrain data, used partial least squares method PLS to select the main wildfire forecasting factors and established an optimized power grid wildfire risk early warning model. Literature [13] designed a forest fire early warning model based on mobile edge computing (MEC) by acquiring ground surface parameters, which can be used to effectively predict wildfires.

In order to more fully combine meteorological data and human factors, this study is based on the MODIS fire point data of a southern province from 2015 to 2019 combined with meteorological data, terrain data, land surface temperature (LST), human factors, and vegetation factors to analyze the influencing factors of mountain fire disasters and establish a CatBoost model to predict fire points. Effective prediction and early warning of fire points are significantly important to reduce the loss of wildfire disasters.

2. Analysis and Data Acquisition of Influencing Factors of Mountain Fire Disasters

The occurrence of mountain fire disasters is comprehensively affected by a variety of factors. According to the analysis of relevant literature and the research on the principles of mountain fires [14], the occurrence of mountain fire disasters is not random, and specific laws have been passed in relation to this situation. This article divides the factors that affect the appearance of wildfires into five aspects: vegetation factors, human factors, surface temperature, terrain factors, and meteorological factors. This research aims to realize large-scale wildfire assessment through multisource remote sensing data and combined meteorological data. The specifically related factors among the five factors that affect the occurrence of wildfire disasters are as follows.

2.1. Remote Sensing Data

2.1.1. Vegetation Factors. Vegetation is the material basis for the occurrence of wildfire disasters. In this study, the influencing factors of plant on wildfire disasters are refined into normalized difference infrared index 7 (NDII7) and normalized differential vegetation index (NDVI). The NDII7 is a critical wildfire risk assessment factor. Qin [15] proved that the NDII7 can characterize the vegetation fuel moisture content and then evaluate the mountain fire risk. The NDVI is used as a criterion for judging surface vegetation and estimate the growth status and density of plant. The occurrence of mountain fire disasters is closely related to the growth status and density of vegetation. Wang et al. [14] judged the event of wildfire disasters and estimated the area of the fired area according to the change of plant NDVI at adjacent time points.

The acquisition of NDVI comes from the MOD13A1 vegetation information product of MODIS provided by the NASA website (<https://ladsweb.modaps.eosdis.nasa.gov/>), with a spatial resolution of 1000 m. The global NDVI information is updated every 16 days. NDII7 is derived from the MOD09A1 product provided on the website as previously mentioned. The temporal resolution of this product is 8 days. After the product is obtained, NDII7 is calculated according to the calculation formula obtained by Qin [15] and others:

$$\text{NDII7} = \frac{\rho_2 - \rho_7}{\rho_2 + \rho_7} \quad (1)$$

2.1.2. Human Factors. The occurrence of wildfire disasters is highly correlated with the time of people's frequent activities. Statistics show that the occurrence of wildfire disasters shows a significant upward trend every Friday and every day from 13:00 to 16:00 from January to April [2]. The uncertainty of human factors is relatively considerable. This study extracts the influencing factors of wildfire disasters as land type, distance from roads, and distance from cultivated land. These data directly indicate the inevitability of human activities and can be used as the influencing factors of wildfire disasters. This notion indirectly suggests the impact of humans on fire. Land types are classified into cultivated land, forest land, grassland, water area, residential land, and unused land according to the 30 m classification data of the global surface.

2.1.3. LST. Surface temperature affects the occurrence of forest fires because it will indirectly affect the moisture content of the combustibles of vegetation. In areas with a relatively dense vegetation, the evaporation of the surface is relatively small because the surface temperature is low, thereby leading to the high moisture content of the combustibles. Mountain fire disasters are less likely to occur [16].

By contrast, if the surface temperature is high, then it is easy to cause mountain fire disasters.

The LST data come from the MOD11A1 product, with a spatial resolution of 1000 m and a temporal resolution of 1 day.

2.1.4. Terrain Factors. Elevation, slope, and aspect are fixed static variables, and many researchers classify them as the fundamental factors leading to wildfire disasters. The ups and downs of terrain will cause different vegetation coverage and meteorological conditions, including rainfall, water content, dense vegetation, vegetation types, and growth conditions; thus, the probability of wildfire disasters will naturally vary. The spatial resolution of terrain data is 30 m. Currently, NASA website (<https://landsweb.modaps.eosdis.nasa.gov/>) provides downloading of SRTM 30 m resolution digital elevation data.

2.2. Meteorological Data. The probability of mountain fire disaster is highly correlated with meteorological factors. Meteorological factors, such as rainfall, average relative humidity, maximum temperature, average temperature, minimum temperature, maximum wind speed, and maximum wind direction [15], have a significant influence on the occurrence of wildfire disasters. The meteorological data come from the China Meteorological Data Network (<http://data.cma.cn/>), which is the cumulative annual value data set (2015–2019) of China.

3. Information Extraction

3.1. Fire Point Information Extraction. The fire point data come from the fire point product of MODIS C6 (2015–2019) provided by <https://firms.modaps.eosdis.nasa.gov/>, and its spatial resolution is 500 m. This study extracts the fire point data according to the fire point collection time and confidence level provided by the product. Detailed MODIS C6 product information is shown in Table 1. This study will extract high-confidence fire data with a confidence of more than 90% as the input data of the fire information to improve the quality of the extracted fire information.

3.2. Nonfire Point Information Extraction. This study first determines the distance of 35 pixels (17,500 m) from the buffer radius of the fire point through the semivariogram function [17] on the basis of the fire point data to eliminate the influence of time and then extracts it from the ring buffer (17,500–18,000 m). Thereafter, all the nonfire point data in a month are obtained. Finally, the daily fire point data corresponding to the fire point data are extracted from the corresponding monthly nonfire point data according to the daily fire point data.

4. Input Data Preprocessing

4.1. Spatial Interpolation of Meteorological Data. The meteorological data downloaded from the China Meteorological Data Network are monitored by various

meteorological stations and are spatially discrete. The meteorological data need to be spatially interpolated to achieve the continuity of the meteorological data in the study area. This study uses Anusplin software to interpolate meteorological data, which has a good effect. Qian et al. [18] compared the interpolation accuracy of Anusplin software with that of Ordinary Kriging and reverse distance weights and found that the interpolation error of the former is the smallest. The interpolation principle is mainly to use ordinary and local thin disk spline functions. The advantage of this method is primarily that it allows the introduction of multiple influence factors as covariates. This study introduces elevation data to significantly reduce the influence of elevation on temperature data changes.

4.2. Data Undersampling. This study will use the ensemble resampling [15] algorithm for undersampling the data to ensure the consistency of the model training samples, that is, the proportion of fire-spot samples and nonfire-spot samples is the same. This algorithm can correctly solve the problem of data loss in the undersampling. Such an algorithm uses ensemble to sample with various models. Each model is undersampling. The undersampling results of multiple models are integrated, and the data distribution will not be changed. The sampling effect is better than the current numerous oversampling and undersampling techniques.

4.3. Normalization of Real Factor Data. Among the influencing factors of mountain fire disasters, some variables are of real number type. Before the CatBoost model is trained, such input data must be normalized to ensure the dimensionlessness of the data, such as the following: distance from the road (x_1 , m), distance from cultivated land (x_2 , m), land surface temperature (x_3), NDVI (x_4), NDII7 (x_5), DEM (x_6 , m), precipitation (x_7 , mm/day), maximum temperature (x_8 , °C), average relative humidity (x_9 , %), average temperature (x_{10} , °C), lowest temperature (x_{11} , °C), and maximum wind speed (x_{12} , m/s). These input variables will be normalized to zero mean. The advantage of this method is that if abnormal points occur, then a small number of strange points will not have a significant effect on the average value; thus, the variance of the variance is little. Z-score normalization is also called standardization. This method maps data to a distribution with a mean of zero and a standard deviation of one. With regard to the above x_i , formula (2) is used to standardize the data, and the obtained new variable data x'_i is used as the input data of the model:

$$x'_i = \frac{x_i - \text{mean}}{\sigma}, \quad (2)$$

where x_i ($i = 1, 2, 3, \dots, 12$) is the original wildfire disaster impact factor, mean; σ is the average value and standard deviation corresponding to each element; and x'_i is the standardized wildfire disaster impact factor.

4.4. Discrete Factor One-Hot Encoding. The discontinuous values, such as land type, slope, and aspect, have no significance. This study will perform one-hot encoding to

TABLE 1: MODIS C6 product details.

Brightness	Scan	Track	ACQ_Date	Satellite	Confidence	Bright_T31	FRP
312.4	1.1	1.1	2015/1/1	Terra	72	290.3	11.6
305.3	1.3	1.1	2015/1/1	Terra	61	289.2	9.1
304.3	1.3	1.1	2015/1/1	Terra	59	292.3	6.9
326.2	1.3	1.1	2015/1/1	Terra	84	289.8	34.1
307.8	1.3	1.1	2015/1/1	Terra	66	288.3	11.2
312	3.3	1.7	2015/1/1	Aqua	72	285.9	67.5
305	4.3	1.9	2015/1/1	Aqua	57	290.3	42.7
304.5	1	1	2015/1/2	Aqua	48	292.2	4.9
324.5	1	1	2015/1/2	Aqua	83	291.6	22.5
309.7	1	1	2015/1/2	Aqua	35	293.6	5.8
346.2	1	1	2015/1/2	Aqua	94	294	55.8
305.6	1	1	2015/1/2	Aqua	62	287.8	6.7
315.4	1	1	2015/1/2	Aqua	73	293.2	12.9
301.1	1	1	2015/1/2	Aqua	44	289.6	3.9
...
300.3	1	1	2019/12/30	Terra	30	286.9	4.3
304.7	1.1	1	2019/12/30	Terra	60	292.1	5.6
308.1	1	1	2019/12/31	Aqua	61	295.5	5
302.9	1	1	2019/12/31	Aqua	24	290.4	3
300.1	1	1	2019/12/31	Aqua	13	285.9	3.8
302.4	1.2	1.1	2019/12/31	Aqua	49	290	5

eliminate the influence between the numerical values. The significant advantage of this method is that it is easy to deal with noncontinuous values, and the model input data are also expanded to a certain extent.

4.5. Feature Selection Method Based on the Gradient Boosting Decision Tree (GBDT) Model. Features must be selected because of the large number of variables in this study, and some variables have little effect on the occurrence of wildfires. Feature selection is the process of choosing factors that are highly correlated with the appearance of fires. The feature selection method based on the GBDT model is a commonly used feature selection method based on the tree model. The principle is to use the node magazines in each decision tree to calculate the importance of features. The final feature importance is the average of the feature importance of all decision trees. In this study, the cross-validation method is used to select the factors whose feature importance is more significant than 0.3. Then, the dimensionality reduction is performed according to the principal component analysis (PCA). The ranking of the importance index of wildfire impact factors is shown in Table 2.

4.6. PCA: Principal Component Analysis. Among the influencing factors of mountain fire disaster, a specific correlation exists between elevation, slope, aspect, maximum temperature, average temperature, minimum temperature, and surface temperature. This study uses the currently widely used linear dimensionality reduction algorithm (PCA) to reduce the dimensionality of all influencing factors of wildfire disasters and eliminate redundant data. The advantage of this algorithm is its ability to retain the original data quality of the sample. In this mechanism, the model training data are compressed as much as possible, and the

factors with high principal components for model training are determined.

The mathematical model of the PCA algorithm in this study is as follows.

$X = x_1, x_2, \dots, x_m$ is the impact factor of wildfire disaster, where the dimension of X is m , which is the number of impact factors. The projection of x_i on the hyperplane in the new hyperdimensional space is $W^T x_i$. The principle is to increase the variance between all sample points to ensure that the projections between all sample data are separated as much as possible. XX^T can be obtained according to the following formula:

$$XX^T W = \lambda W. \quad (3)$$

After the sample feature matrix XX^T is decomposed, the eigenvalues of each factor $\lambda_1 \geq \lambda_2 \geq \dots \lambda_m$ are obtained, and the corresponding eigenvectors of the first I samples $W = (W_1, W_2, \dots, W_i)$ are the required mountain fires of the principal components of the disaster impact factor. This paper retains 99% of the main information of the original feature. The latitude of the principal component m is 18. Compared with the feature selection based on the GBDT model, the feature dimension is reduced by 13.

5. Fire Point Prediction Model Based on CatBoost Algorithm

5.1. CatBoost Model. CatBoost is an algorithm that combines GBDT and categorical features. This approach is an improved implementation under the framework of the GBDT algorithm. CatBoost is based on oblivious trees with few parameters and supports categorical variables and high accuracy sexual GBDT framework. The main pain point is to efficiently and rationally deal with categorical features. CatBoost is composed of categorical variables and boost.

TABLE 2: Importance index of wildfire impact factors.

Land surface temperature	37.61
Distance from the road	7.31
Maximum wind speed	6.51
Slope	5.15
Aspect (NE)	0.63
Aspect (SW)	0.58
Aspect (N)	0.54
Precipitation	0.37
Wind direction (SW)	0.29
Aspect (E)	0.16
Wind direction (SE)	0.07
Wind direction (NNE)	0.00
Wind direction (ENE)	0.00
Normalized infrared index 7	9.83
Maximum temperature	6.76
Normalized vegetation index	6.24
Wind direction (WSW)	0.67
Aspect (N)	0.62
Aspect (NW)	0.55
Grass	0.44
Aspect (W)	0.35
Wind direction (SSE)	0.28
Wind direction (W)	0.16
Wind direction (E)	0.06
Wind direction (NNW)	0.00
Wind direction (N)	0.00
Elevation	9.69
Distance from cultivated land	6.63
Average relative humidity	5.84
Wind direction (S)	0.65
Aspect (SE)	0.62
Woodland	0.55
Cultivable land	0.42
Wind direction (WNW)	0.32
Wind direction (SSW)	0.27
Wind direction (ESE)	0.12
Wind direction (NW)	0.04
Wind direction (NE)	0.00
Aspect (flag)	0.00

This mechanism also deals with gradient bias and prediction shift problems, thereby improving the generalization ability and robustness of the algorithm [19, 20]. This study considers many categorical features, such as rainfall, wind direction, slope direction, and land type. CatBoost can be used to quickly process nonnumerical features. When the CatBoost algorithm processes categorical features, it puts all sample data sets into the algorithm for learning. Then, CatBoost randomly arranges all these sample data sets and filters out samples with the same category from all features. When numerically transforming the characteristics of each sample, the target value of the sample is first calculated before the sample, and the corresponding weight and priority are added [21, 22]. The specific formula is shown in the following:

$$x_k^i = \frac{\sum_{j=1}^n \{x_j^i = x_k^i\} \cdot y_i + ap}{\sum_{j=1}^n \{x_j^k = x_k^i\} + a}, \quad (4)$$

where p represents the added prior value and the weight coefficient greater than zero. An a priori value is added to significantly reduce the noise points caused by low-frequency features to effectively minimize the overfitting of the model and improve the generalization ability.

5.2. Fire Point Model Training and Optimization

5.2.1. Model Training. The five-year MODIS monitoring fire point data of a southern province from 2015 to 2019 and the nonfire point data extracted by the method described in this study are selected as the sample set. The fire-spot data with a confidence level of less than 90% is eliminated to improve the quality of the fire-spot samples. The sample data after data oversampling, normalization, one-hot encoding, feature selection, and PCA dimensionality reduction are substituted into the CatBoost model for training. Approximately 70% of the data are randomly selected for model training and 30% for model testing. The temporal resolution of NDII7, NDVI, and land surface temperature in the input feature variables of the model are 8 days, 16 days, and 1 day, respectively. The input data of NDII7, NDVI, and land surface temperature select the data of the previous time phase before the fire to prevent the input vegetation data and land surface temperature from being affected by the fire and failing to achieve the effect of fire prediction. The data of human factors and terrain factors are unchanged, while the input time phase of weather data is consistent with the fire data. Figure 1 is the time phase relationship diagram of the input feature variables of the CatBoost model, and Figure 2 is the basic flow chart of model training.

5.2.2. Model Optimization. This study uses grid search combined with tenfold cross-validation to optimize the primary hyperparameters of the CatBoost model, including iterations, learning_rate, max_depth, criterion, and feature importance, to improve the accuracy of model fire prediction. Tenfold cross-validation divides the sample data into ten mutually exclusive training subsets. Each time nine subsets are selected as training data, and the remaining subset is used as test data. The multiple rounds of training are repeated to ensure that each subset is as the test set, the ten test results are obtained, and the average of the ten test results is the accuracy of the model. The hyperparameters obtained through a grid search can effectively improve the prediction effect of the model [23].

After model optimization, the best hyperparameters of the fire point prediction model are shown in Table 3.

5.2.3. Model Evaluation. This study uses accuracy, precision, recall, F1-score, and AUC value to make a comprehensive evaluation of the model prediction accuracy and address the classification problem of unbalanced data of fire point prediction. The confusion matrix of the fire point and nonfire point data sets in this article is shown in Table 4.

The evaluation index of the fire point prediction model can be obtained according to the confusion matrix.

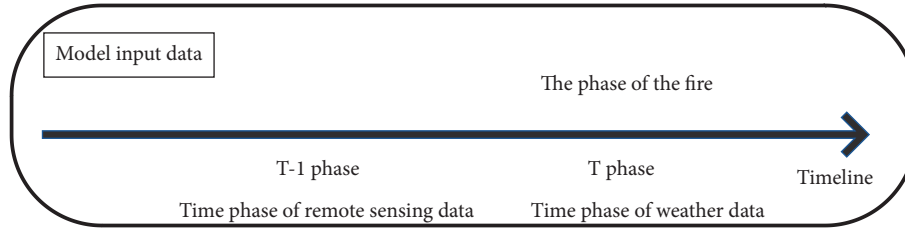


FIGURE 1: Time-phase relationship diagram of the model input characteristic variables.

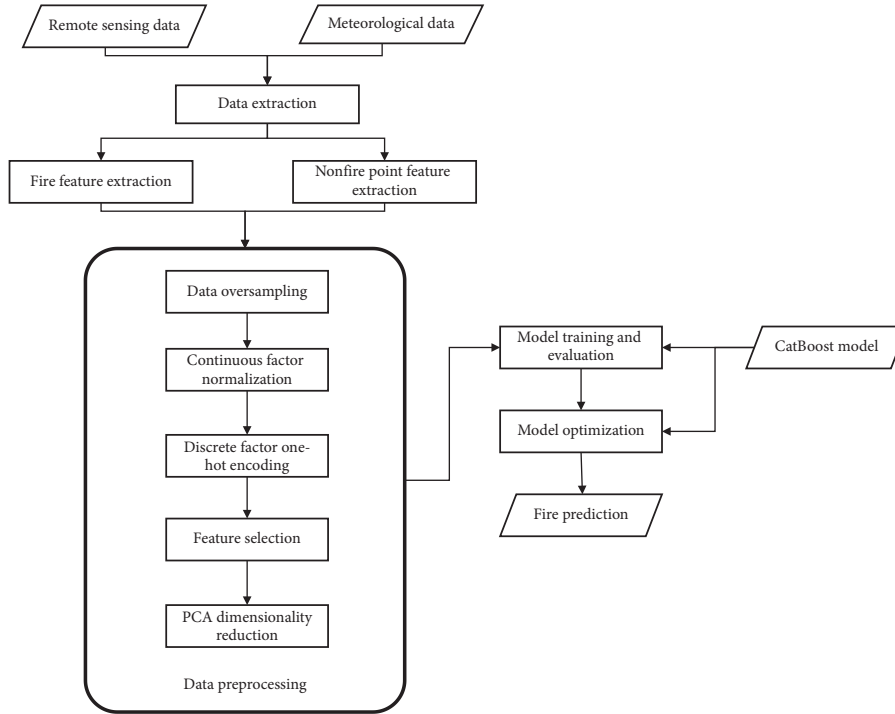


FIGURE 2: Flow chart of the fire prediction model.

TABLE 3: Best hyperparameters of the fire prediction model.

Iterations	Learning_rate	Max_depth	Criterion	Feature importance
1000	0.05	9	Gini	0.3

TABLE 4: Confusion matrix of fire and nonfire points.

	Predicted fire point	Predicted nonfire point
Real fire point	TN (true negative)	FP (false positive)
Real nonfire point	FN (false negative)	TP (true positive)

TN: the actual value is the fire point, and it is also predicted as the fire point. FN: the actual value is a nonfire point, but it is predicted to be a fire point. FP: the actual value is a fire point, but it is predicted to be a nonfire point. TP: the actual value is a nonfire point, and it is also predicted as a nonfire point.

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FN + FP + TN)}. \quad (5)$$

$$\text{Recall} = \frac{TP}{(TP + FN)}. \quad (7)$$

$$\text{Precision} = \frac{TP}{(TP + FP)}. \quad (6)$$

$$\text{F1-score} = 2 \frac{(\text{precision} * \text{recall})}{(\text{precision} + \text{recall})}. \quad (8)$$

TABLE 5: CatBoost fire point prediction model results.

	Precision	Recall	F1-score	Support	AUC
Nonfire point	0.83	0.87	0.78	889	
Fire point	0.81	0.82	0.83	273	
Accuracy			0.79	1162	
Overall	0.82	0.84	0.80	1162	0.83

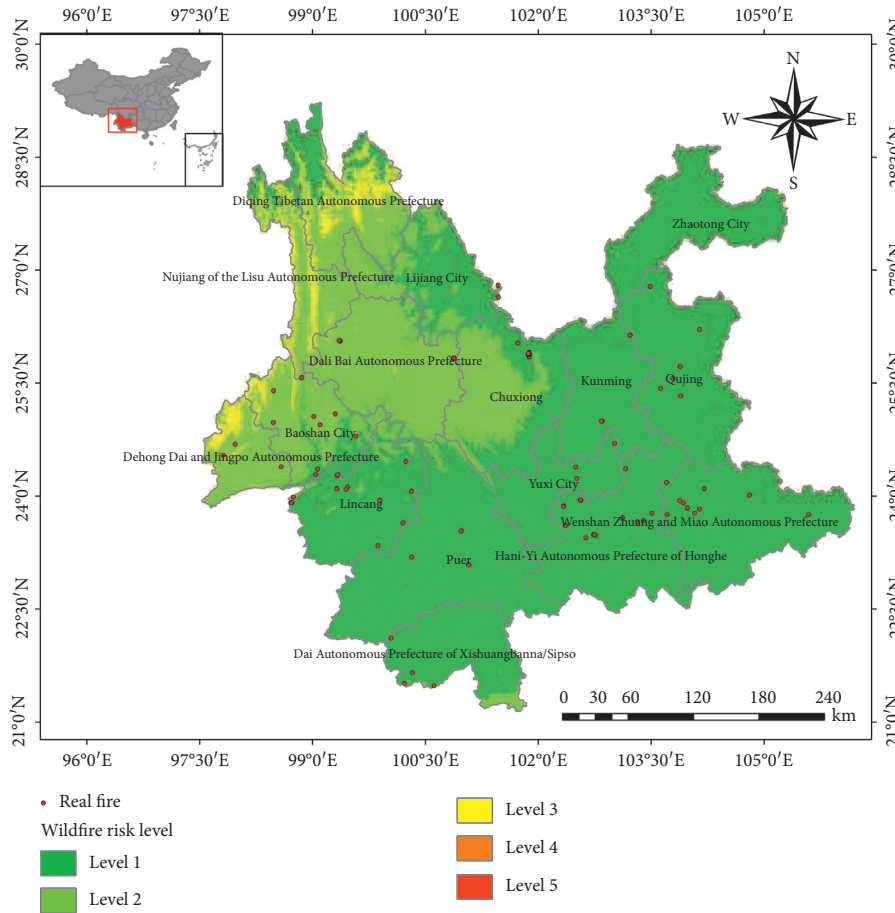


FIGURE 3: The risk forecast of wildfires in Yunnan Province on March 15, 2020.

AUC value: the AUC value is the area value under the ROC curve, which can quantitatively reflect the model performance measured on the basis of the ROC curve. The abscissa of the ROC curve is the false positive rate, $FPR = FP / (FP + TN)$, and the ordinate is the true positive rate, $TPR = TP / (TP + FN)$.

This study uses the best hyperparameters obtained from the model optimization in Section 4.2 to predict the fire point of the sample data set. The final five model evaluation indicators are shown in Table 5.

The results shown in Table 5 demonstrate that the CatBoost fire point prediction model after model optimization has a nonfire point precision of 0.83, recall of 0.87, and F1-score of 0.78 and a fire point precision of 0.81, recall of 0.82, and F1-

score of 0.83. The final accuracy is 0.79, the overall precision is 0.82, recall is 0.84, the F1-score is 0.80, and the AUC value is 0.83. The fire prediction results indicated that the model's prediction of the fire starts with a good predictive effect, and the risk of wildfires can be effectively predicted.

In order to more intuitively reflect the effect of the model in predicting the risk of wildfires, this article draws the comparison between the wildfire risk prediction maps and real fire spots in Yunnan Province on March 15, 2020, April 15, 2020, and May 15, 2020. The resolution of the wildfire risk prediction map is 500 meters, as shown in Figures 3–5. It can be seen that more than 80% of the real fire points fall in the high-risk area of the prediction map, which further verifies the model's effectiveness.

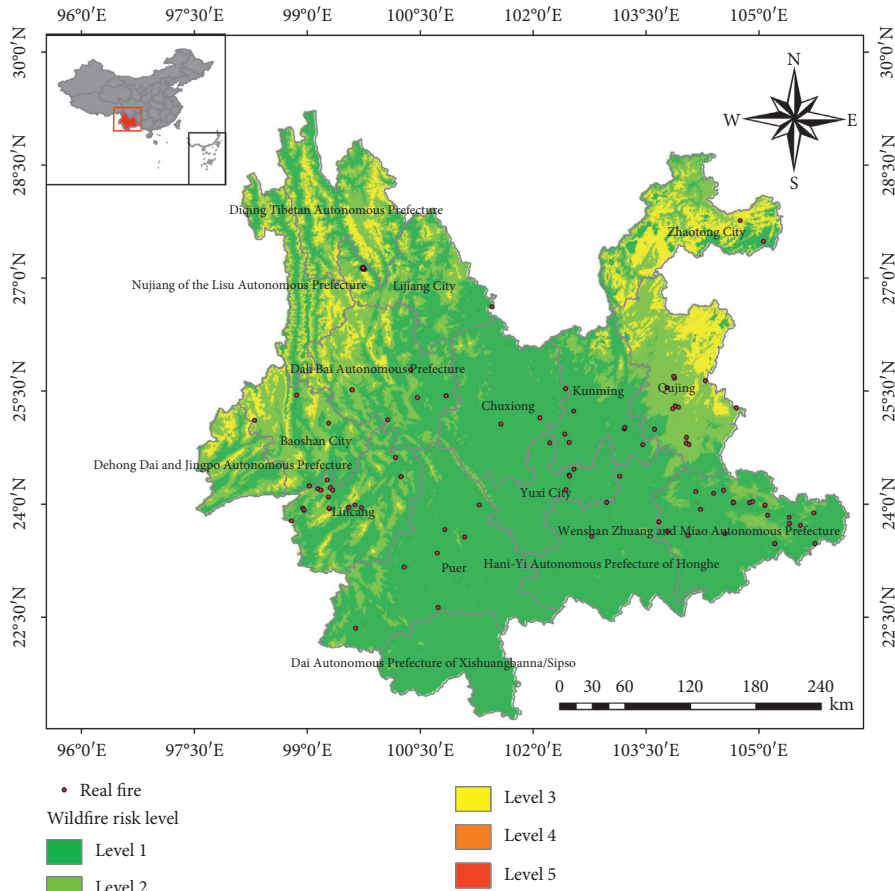


FIGURE 4: The risk forecast of wildfires in Yunnan Province on April 15, 2020.

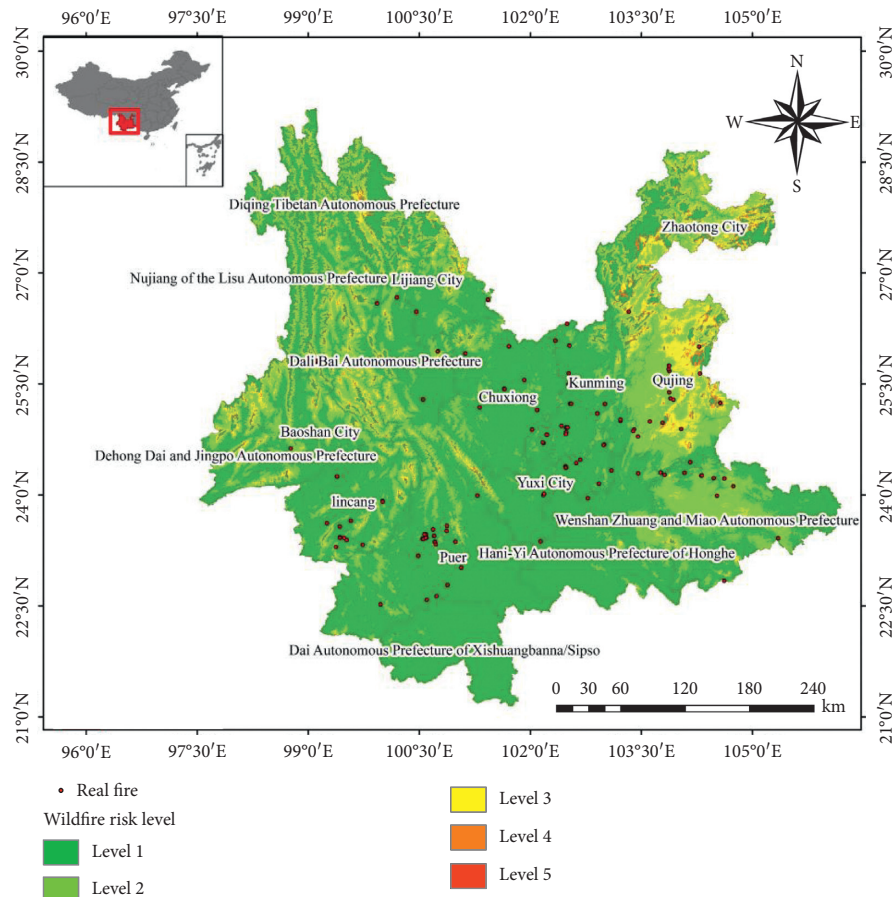


FIGURE 5: The risk forecast of wildfires in Yunnan Province on May 15, 2020.

6. Conclusions and Prospects

This study uses MODIS fire data, combined with vegetation factors, human factors, meteorological factors, surface temperature, and terrain factors, based on feature selection and PCA dimension reduction to find out the influencing factors that are highly correlated with the occurrence of wildfires. The research proposes a method based on CatBoost algorithmic fire prediction model. This model can effectively predict fire points, is helpful in preventing wildfire risks, and has a specific guiding role for the electric power department to avoid risks of fire and make appropriate early warning arrangements in advance.

Although this article has achieved positive research results, it still has some deficiencies and areas worthy of in-depth study. The research conducted in this study is only based on the first-level classification of land types to make fire forecasts and does not make precise fire forecasts under a single ground type. Under the secondary classification of land types, the establishment of different fire prediction models is based on each specific feature to achieve more precise and accurate fire prediction in the direction of further in-depth research.

Data Availability

This article contains data to support the results of this research. Some data cannot be provided because it involves the coordinate data of power grid poles.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

- [1] L. Shun-gen, "Forest-fire rule, cause and preventive measure," *China New Technologies and Products*, vol. 3, no. 252, pp. 354-355, 2011.
- [2] X. Chen, R. N. Ling, J. Chen et al., "A temporal and spatial clustering analysis and its application with wildfire disasters in hubei power grid based on satellite hot spot data," *Bulletin of Surveying and Mapping*, vol. 268, no. 9, pp. 14-18, 2015.
- [3] G. Lei, W. He, and J. Lin, "Realization and application of integrated monitoring system for anti-landscaping in overhead transmission line," *Electrical Engineering*, vol. 12, no. 12, pp. 112-115, 2013.
- [4] H. Xu, W. Liao, K. Li, and T. Guo, "Measures for preventing mountain fire in transmission line," *Electric Safety Technology*, vol. 5, no. 5, pp. 51-54, 2013.
- [5] M. Liu, Z. Shao, S. Guan et al., "Study on models and methods of risk assessment of transmission line fault caused by fire," *Power System Protection and Control*, vol. 43, no. 15, pp. 28-35, 2015.
- [6] X. Qin, Z. Zhang, and Z. Li, "Forecasting methodology of national-level forest fire risk rating," *Remote Sensing Technology and Application*, vol. 486, no. 5, pp. 500-504, 2008.
- [7] Y. Wang, H. Cheng, and W. Duan, "A new forecast method of forest fire risk rating based on remote sensing data of land surface temperature and humidity," *Journal of West China Forestry Science*, vol. 43, no. 6, pp. 97-103, 2014.
- [8] E. Zhou, Y. Huang, J. Chen, X. Tian, R. Wei, and Y. Zhou, "Graph theory based fire risk prediction level model of overhead transmission lines," *Southern Power System Technology*, vol. 4, no. 4, 2020.
- [9] X. Xiong, Y. Zeng, J. Wang, and H. Li, "Risk assessment of power transmission channels in forest regions based on spatial-temporal features of forest fire," *Power System Protection and Control*, vol. 46, no. 4, pp. 1-9, 2008.
- [10] Z. Feng, Y. Lu, Y. Sun, Q. Xu, and P. Jiang, "Assessment of power grid risk caused by wildfire disaster in Yunnan Province," *Journal of Natural Disasters*, vol. 23, no. 5, 2014.
- [11] J. Lu, J. Guo, Li Yang, F. Tao, and J. Zhang, "Research and application of fire forecasting model for electric transmission lines incorporating meteorological data and human activities," *Mathematical Problems in Engineering*, vol. 2016, Article ID 9828676, 10 pages, 2016.
- [12] E. Zhou, S. Hu, L. Zhang, R. Wei, H. Wang, and Y. Fan, "Characteristics and risk warning technology of wildfire disaster in power grid," *Electric Power Engineering Technology*, vol. 39, no. 3, pp. 2096-3203, 2020.
- [13] C. Cheng, H. Zhou, X. Chai et al., "Adoption of image surface parameters under moving edge computing in the construction of mountain fire warning method," *PLoS One*, vol. 15, no. 5, p. e0232433, 2020.
- [14] H. Wang, J. Luo, T. Xu et al., "Investigation and analysis of the status quo of natural disaster prevention technology in China power grid," *Automation of Electric Power Systems*, vol. 34, no. 23, pp. 5-10, 2010.
- [15] X. Qin, "Study on forest fire early warning and monitoring methodology using remote sensing and geography information system techniques," PhD thesis, Resource Information Institute of Chinese Academy of Forestry, Chinese Academy of Forestry, Beijing China, 2005.
- [16] H. K. Preisler, R. E. Burgan, J. C. Eidenshink, J. M. Klaver, and R. W. Klaver, "Forecasting distributions of large federal-lands fires utilizing satellite and gridded weather information," *International Journal of Wildland Fire*, vol. 18, no. 5, pp. 508-516, 2009.
- [17] R. Fraser, Z. Li, and J. Cihlar, "Hotspot and NDVI differencing synergy (HANDS) A new technique for burned area mapping over boreal forest," *Remote Sensing of Environment*, vol. 74, no. 3, pp. 362-376, 2000.
- [18] Y. Qian, H. Lv, Y. Zhang et al., "Application and assessment of spatial interpolation method on daily meteorological elements based on ANUSPLIN software," *Journal of Meteorology and Environment*, vol. 26, no. 2, 2010.
- [19] B.-h. Huang, H. Zhang, Z.-j. Sun et al., "Forest fire danger factors and their division in Shandong based on GIS and RS," *Chinese Journal of Ecology*, vol. 34, no. 5, pp. 1464-1472, 2015.
- [20] A. L. F. Chan, "An application of classification analysis for skewed class distribution in therapeutic drug monitoring-the case of vancomycin," in Proceedings of the 2004 IDEAS Workshop on Medical Information Systems: The Digital Hospital (IDEAS-H'04), Beijing, China, September 2004.
- [21] Dorogush A. V., Ershov V., Gulin A.. CatBoost: Gradient Boosting with Categorical Features Support [EB/OL]. (2018-11-24). [2019-03-19], <https://arxiv.org/abs/1810.11363>.
- [22] Prokhorenkova L., Gusev G., Vorobev A., et al.. CatBoost: Unbiased Boosting with Categorical Features [EB/OL]. (2019-01-02). [2019-03-22], <https://arxiv.org/abs/1706.09516>.
- [23] J. Wang, L. Zhang, G. Chen et al., "A parameter optimization method for an SVM based on improved grid search algorithm," *Applied Science and Technology*, vol. 391, no. 3, pp. 28-31, 2012.

Research Article

Prediction of Financial Time Series Based on LSTM Using Wavelet Transform and Singular Spectrum Analysis

Qi Tang ¹, Ruchen Shi ¹, Tongmei Fan ¹, Yidan Ma ² and Jingyan Huang ¹

¹School of Economics and Management, Southeast University, Nanjing, China

²School of Cyber Science and Engineering, Southeast University, Nanjing, China

Correspondence should be addressed to Qi Tang; regantangqi@seu.edu.cn

Received 11 March 2021; Revised 6 May 2021; Accepted 23 May 2021; Published 9 June 2021

Academic Editor: Jie Chen

Copyright © 2021 Qi Tang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In order to further overcome the difficulties of the existing models in dealing with the nonstationary and nonlinear characteristics of high-frequency financial time series data, especially their weak generalization ability, this paper proposes an ensemble method based on data denoising methods, including the wavelet transform (WT) and singular spectrum analysis (SSA), and long-term short-term memory neural network (LSTM) to build a data prediction model. The financial time series is decomposed and reconstructed by WT and SSA to denoise. Under the condition of denoising, the smooth sequence with effective information is reconstructed. The smoothing sequence is introduced into LSTM and the predicted value is obtained. With the Dow Jones industrial average index (DJIA) as the research object, the closing price of the DJIA every five minutes is divided into short term (1 hour), medium term (3 hours), and long term (6 hours), respectively. Based on root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and absolute percentage error standard deviation (SDAPE), the experimental results show that in the short term, medium term, and long term, data denoising can greatly improve the stability of the prediction and can effectively improve the generalization ability of LSTM prediction model. As WT and SSA can extract useful information from the original sequence and avoid overfitting, the hybrid model can better grasp the sequence pattern of the closing price of the DJIA.

1. Introduction

As the worldwide largest economy, the US has advanced statistical constitutions and a mature financial supervision system, whose financial data are comprehensive, accurate, and credible. At the same time, the US stock market cooperates with other markets in an efficient way and plays an important role in the US financial system, and all these characteristics make the market a good model. On the one hand, global stock markets react quickly to the tendency of this market, especially in the case of unusually high market volatility. On the other hand, most economic theories and assumptions are based on the study of a developed financial system with a larger and more active stock market, a more mature economy, and a more effective financial supervision system. As a representative of developed markets, the US market is also the most favorable object for empirical or

theoretical propositions in academic research. The three major stock indexes in the United States are the Dow Jones Index [1], the Standard and Poor's 500 Index, and the Nasdaq Composite Index. The most famous of these indexes is the Dow Jones Index. The importance of the Dow Jones Index has been further recognized in global markets beyond its role in the domestic market. The 30 companies that make up the Dow, such as Citigroup, Coca-Cola, General Motors, and Intel, are prestigious multinational corporations. They cover a wide range of large industries and their performance is behind the global economy. Therefore, forecasting the Dow Jones index is of great significance to the entire financial system.

At present, there are two categories of prediction models that are suitable for financial time series: parametric model and nonparametric model. Autoregressive (AR), moving average (MA), autoregressive moving average (ARMA), and

autoregressive comprehensive moving average (ARIMA) are typical models of parameter types [2]. However, all of the above models can only be used if the predicted time series conforms to the statistical assumptions they have. Therefore, the parametric model has limitations to some degree. Additionally, the parametric model is more functional for the time series with linear characteristics, but the financial time series of DJIA price is characterized with nonlinearity and is highly fluctuating, which will cause so many estimated parameters and increase the complexity of the model. Due to the limitations of the parametric model and features of DJIA price, the parametric model is unsuitable for the Dow Jones index price forecast.

Stock price prediction is a pertinent and tricky problem that has caught the interest of many scholars. In order to adapt to the characteristics of DJIA and ensure high accuracy, nonparametric models are used. Specifically, machine learning (ML) and deep learning (DL) models are used for DJIA prediction. Ahmed [3] found that machine learning methods can perform better than traditional econometrics methods. Sun et al. [4] have compared the accuracy of the echo state network (ESN) and the long short-term memory model (LSTM) on Kweichow Moutai's stock price prediction. The empirical results found that the ESN model can alleviate the problems of the low accuracy of the deep learning model, the slow convergence speed, and the complex network structure. At the same time, it has higher prediction accuracy. Cao et al. [5] used the convolutional neural network (CNN) model to predict the stock index and finally established the CNN-support vector machine (SVM) stock index prediction model. The empirical results show that the effect of using a neural network to predict the financial time series is better than the traditional measurement method. Ustali et al. [6] have used Artificial Neural Network (ANN), Random Forest (RF) algorithm, and XGBoost algorithm to estimate the future price of company stocks listed on the Istanbul Stock Exchange (BIST) United Joint Stock Company (BIST) 30 Index. The empirical results show that although the results of XGBoost and Random Forest algorithms are similar, the prediction results of XGBoost are slightly better. Moreover, the performance of both models is better than ANN.

LSTM (Long Short-Term Memory) network is one of the cyclic neural networks (RNNs). Sepp Hochreiter and Jurgen Schmidhuber first presented this algorithm in Neural Computing [7]. It has better performance than ordinary RNN in data processing and prediction. Considering the excellent performance of the LSTM network in time series, Jiang et al. [8], taking the daily data of Shanghai Composite Index and Dow Jones Index as the research object, respectively, uses RNNs and LSTM to build the model. Then they found through experiments that the LSTM model prevails over the RNNs model for the neural network model. However, this model is still not that well applicable to Dow Jones Index. Actually, to ensure the prediction accuracy of the Dow Jones Industrial Average, more potential factors, such as the influence of policy information, should be taken into consideration. However, the quantification of these parameters is exceedingly difficult. Therefore, another

method is to further mine the information of the sequence itself through the lag term.

In order to make up for the deficiency of a single forecasting model, a new forecasting model-hybrid model is presented. In this category, decomposition methods, such as empirical mode decomposition (EMD) [9] and singular spectrum analysis (SSA) [10], are usually combined with ML and DL based on financial time series prediction models [11–13]. In recent works [14, 15], the superiority of hybrid models has been verified because of their preponderance in identifying time series patterns. As there is a nonlinear relationship between the predicted price of agricultural products and the influencing factors, Jia et al. [16] have designed a neural network model of LSTM-DA (Long Short-Term Memory-Double Attention) which combines the convolutional attention network, the LSTM network, and the attention mechanism. Compared with the traditional signal model, this model can improve the prediction accuracy, and the predicted price index can accurately describe the overall trend of vegetable products in the next week.

However, hybrid models are rarely used to forecast DJIA prices. Meanwhile, lots of previous studies have proved that denoising the high-frequency time series can significantly raise the ability to extend the model and dramatically optimize the prediction results. At present, the empirical data decomposition and noise reduction methods mainly include ensemble empirical mode decomposition (EEMD), singular spectrum analysis (SSA), and wavelet transform decomposition (WT). Although the integrated empirical mode decomposition (EEMD) can suppress the mode aliasing problem to some extent, it may increase the complexity of the sequence. Jung et al. [17] have integrated wavelet transforms and recurrent neural network (RNN) based on artificial bee colony (ABC) algorithm (called ABC-RNN) to establish a system for the purpose of stock price prediction, and it turns out that the performance of the presented model is the best in TAIEX. However, it still has some insufficiencies. For example, this system lacks a solution that includes a feature selection function and addition parameter information provision function to achieve a simplified system organization.

This paper proposes an integrated method to establish a prediction model, which utilizes data denoising methods including the wavelet transform (WT) and singular spectrum analysis (SSA), and long-term short-term memory neural network (LSTM). At the same time, we compare the results of different models. The second part of this paper will introduce the model formula used in this paper, and the third part will introduce the data, model prediction accuracy, and stability results comparison. The fourth part is a summary.

2. Model Formulation

In this section, based on the previous work [18], we provide an overview of the main models used in this study, including the LSTM, WT, hybrid WT-LSTM, SSA, and hybrid SSA-LSTM models.

2.1. Long Short-Term Memory. LSTM neural network was first proposed by Hochreiter and Schmidhuber, which is widely used to process sequence information owing to its advantages in discovering long-term dependencies. Therefore, it is theoretically feasible to establish an LSTM neural network model for financial high-frequency time series data. The structure of each neuron in LSTM is shown in Figure 1, and its internal structure is constituted of a cell and three gates. Cell records the state of neurons, and the function of input gate and output gate is to receive, output, and modify parameters. Forget gate controls the forgotten degree of the previous state of neurons; that is, it determines the information to be removed from the cell. The selection of activation function is an important part in the process of training a neural network, which can make the neural network learn the nonlinear factors in the data. The activation function used in this paper is the traditional sigmoid activation function and tanh activation function. It mainly includes the following stages.

The first stage is eliminating part of the information from the cell through the forget gate.

$$f_t = \sigma(b_f + W_f x_t + U_f h_{t-1}), \quad (1)$$

where x_t represents the input of the current cell, h_{t-1} represents the output of the previous cell, and σ represents the sigmoid activation function which reads the information of x_t and h_{t-1} and outputs a value between 0 and 1.

The second stage is updating the status of information in the cell.

$$\begin{aligned} i_t &= \sigma(b_i + W_i x_t + U_i h_{t-1}), \\ \tilde{C}_t &= \tanh(b_c + W_c x_t + U_c h_{t-1}), \\ C_t &= f_t \times C_{t-1} + i_t \times \tilde{C}_t, \end{aligned} \quad (2)$$

where C_{t-1} indicates the status of old cell information.

The third stage is outputting the information controlled by output gate. Firstly, we run the sigmoid layer to determine which part of the cell state outputs. Secondly, we process the cell state through the tanh activation function (get a value between -1 and 1) and multiply it with the output of the sigmoid layer. In the end, we will only output the part of the controlled information.

$$\begin{aligned} O_t &= \sigma(b_o + W_o x_t + U_o h_{t-1}), \\ h_t &= O_t \times \tanh(C_t). \end{aligned} \quad (3)$$

For the purpose of verifying the effectiveness and versatility of WT and SSA filtering scheme, the most common LSTM neural network structure is adopted in this research. The characteristic quantity is selected as the most basic highest price in the last five minutes, lowest price in the last five minutes, closing price in the last five minutes, and opening price in the last five minutes. Specifically, the main structure of the LSTM neural network in this paper includes a 150-node LSTM layer, a 50-node LSTM layer, and a fully connected layer. Besides, a dropout layer is introduced to this model in order to compare the performance of the dropout layer and data

denoising methods. The calculation diagram structure of the LSTM neural network constructed in this paper is shown in Figure 2. The dotted box represents the neural network structure.

2.2. Wavelet Transform Analysis. Dow Jones Index is susceptible to a large number of factors such as economic development, policy changes, and investor sentiment. They usually contain lots of noise and are characterized by nonlinearity. To raise the ability to extend of the model, the noise data should be filtered out when the deep neural network is used to process the nonlinear data. Wavelet analysis can carry out multiscale refined analysis of signals through operation functions such as stretching and shifting, effectively eliminating noise contained in the data and retaining the characteristics of original signals [19]. Accordingly, this paper intends to use wavelet decomposition and reconstruction for data preprocessing of financial time series, which is shown in Figure 3, and adopts "wavelet denoising" to eliminate the high-frequency components of noise in time series, so as to weaken the influence of short-term noise disturbance on neural network structure and improve the prediction performance of the model.

Wavelet decomposition decomposes each input signal into a signal of low frequency and a signal of high frequency and merely decomposes the part of low frequency. Assume that C_0 is the original financial time series signal; $C_1, C_2 \dots C_l$ and $D_1, D_2 \dots D_l$ are the first, second, and L-layer low-frequency and high-frequency signals. Then it can be mathematically expressed as follows:

$$C_0 = C_l + D_l + D_{l-1} + \dots + D_2 + D_1 \quad (4)$$

In order to denoise, the Mallat wavelet is used to be reconstructed on the basis of the coefficients of the first N layer with low frequency and the coefficients of $1-N$ layer with high frequency of wavelet decomposition, and the high-frequency part of zero. The low frequency part of wavelet decomposition of financial sequential data reports the general tendency of the series, and the high-frequency part reports the short-term stochastic disturbance of this financial time series. Therefore, on the one hand, setting the high-frequency part to zero can eliminate the noise and smooth the signal. On the other hand, it can also obtain the approximate signal of the primary financial sequential data so as to prevent the excessive learning of the neural network structure caused by short-term stochastic disturbance factors and raise the extrapolation and generalization ability of the model.

2.3. WT-LSTM. In order to obtain higher accuracy, LSTM and WT are combined to predict the price of DJIA. Hybrid WT-LSTM consists of the following three phases and the modeling process is shown in Figure 4.

The first stage is WT decomposition: sym wavelet is chosen as the wavelet basis for DJIA closing price forward fractional solution.

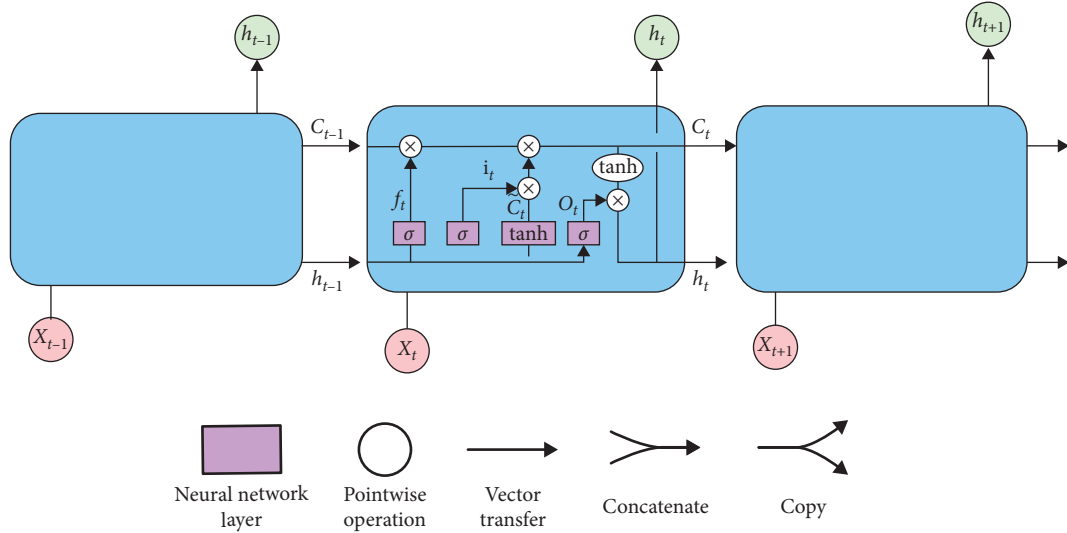


FIGURE 1: LSTM neuron structure.

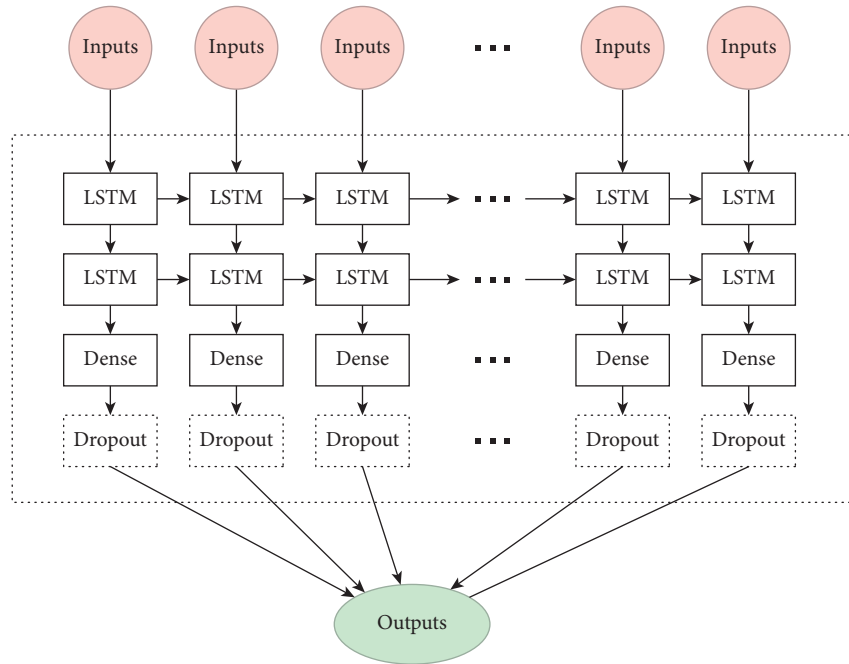


FIGURE 2: LSTM deep neural network structure.

The second stage is time series reconstruction: reconstruct the financial time series data and raise the ability to extend the prediction model.

The third stage is LSTM prediction: smoothed series \tilde{x}_t and volume (V_t) are the input characteristics of LSTM. According to the partial autocorrelation function (PACF) of x_t , the time lag of \tilde{x}_t and V_t is determined. Then, the final prediction result is obtained.

The sym wavelet is an approximate symmetric orthogonal wavelet function of db wavelet, and it has better symmetry [20]. The more the layers of wavelet decomposition, the better the stability of detail signal and approximate signal. However, it will lead to greater errors in

the decomposition process, so the number of layers should not be too much or too little. Under the circumstance of four layers of decomposition, the effect of denoising is remarkable without eliminating much valid information. Therefore, in this paper, the sym4 wavelet basis is firstly used to divide the closing price of DJIA into four layers, so as to reconstruct the time series data of the gold melting and raise the ability to extend the prediction model. On this basis, the general tendency and market volatility information in the primary data are pre-processed. As a result, the hybrid WT-LSTM model proposed can avoid overfitting and outperform the single LSTM model.

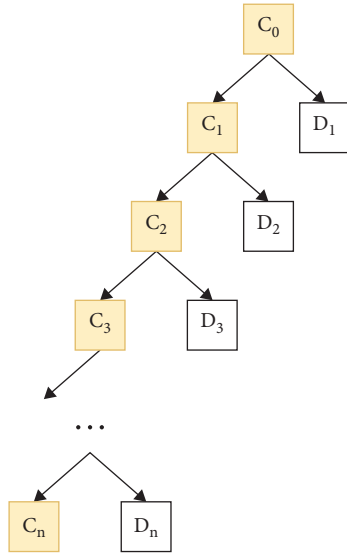


FIGURE 3: Financial time series wavelet decomposition.

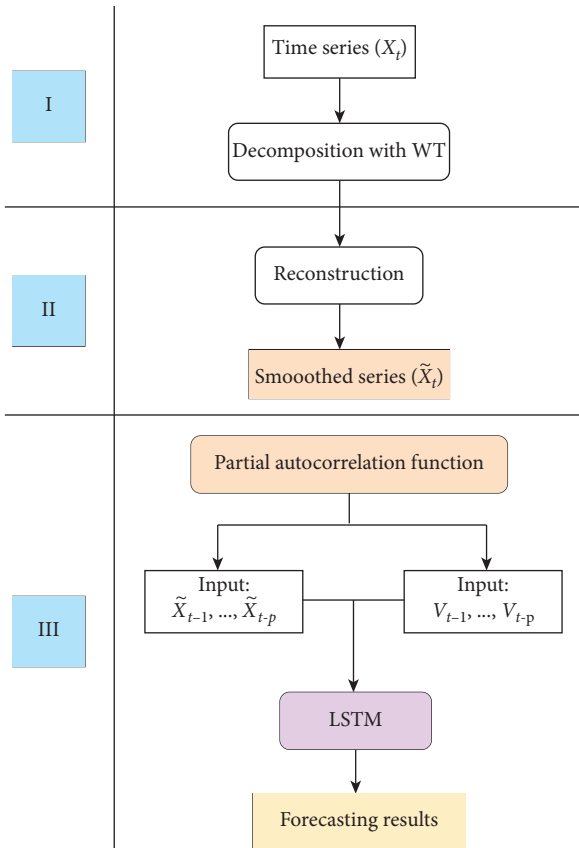


FIGURE 4: Hybrid WT-LSTM model processing process.

2.4. *Singular Spectrum Analysis and SSA-LSTM.* In order to study the effects of the prediction accuracy of out-of-sample data and the prediction ability of future dynamic trend after decomposing and reconstructing financial sequential data by wavelet transform (WT), we also carry out a controlled experiment to study the model prediction after financial time series data decomposition and

reconstruction using singular spectrum analysis (SSA). The effect of SSA can construct a trajectory matrix from the observed financial time series, decompose and reconstruct the trajectory matrix, and extract the different parts of the signal, thus effectively eliminating the noise of the financial sequential data and retaining the features of the primary signal.

Suppose there is a one-dimensional sequence $x(i) (i = 1, 2, \dots, n)$. Given that the embedding dimension is $m (m < (n/2))$, a time-delay matrix X can be obtained, and its dimension is $m \times k (k = n - m + 1)$,

$$X = \begin{bmatrix} x_1 & \cdots & x_{n-m+1} \\ \vdots & \ddots & \vdots \\ x_m & \cdots & x_n \end{bmatrix}. \quad (5)$$

Let S be the $m \times m$ dimensional covariance matrix of the delay matrix, then

$$S = \begin{bmatrix} s(0) & \cdots & s(m-1) \\ \vdots & \ddots & \vdots \\ s(m-1) & \cdots & s(0) \end{bmatrix}. \quad (6)$$

Singular spectrum analysis is used to decompose the covariance matrix S to obtain m singular values $\lambda_i (i = 1, 2, \dots, m)$. Then arrange the obtained m singular values in descending order. The magnitude of the singular value represents the relative relationship between the signal and the noise. The singular value points with larger values are regarded as signal points, and the points with smaller values are regarded as noise points. The eigenvector E^k corresponding to λ_k is called the empirical orthogonal projection function. The orthogonal projection coefficient of the sampled signal $x(i)$ on the eigenvector E^k is the k -th principal component:

$$a_i^k = \sum_{j=1}^m x_{i+j} E_j^k \quad (0 \leq i \leq n - m). \quad (7)$$

If each principal component and the empirical orthogonal function are known, the process of reversing the original sequence is as follows:

$$x_{i+j} = \sum_{j=1}^m a_i^k E_j^k \quad (1 \leq j \leq m). \quad (8)$$

So as to reach a comparison model with the WT-LSTM model, we combined LSTM and SSA to predict the DJIA. The hybrid SSA-LSTM consists of the following three stages and the modeling process is shown in Figure 5.

The first stage is SSA decomposition: SSA technique is used to decompose the primary sequential data (x_t) into tendency, market volatility, and noise.

The second stage is time series reconstruction: reconstruct the smooth sequence based on the tendency and market volatility signal (\tilde{x}_t).

The third stage is LSTM prediction: smooth series \tilde{x}_t and volume (V_t) are the input characteristics of LSTM. In light of the partial autocorrelation function (PACF) of x_t , the time

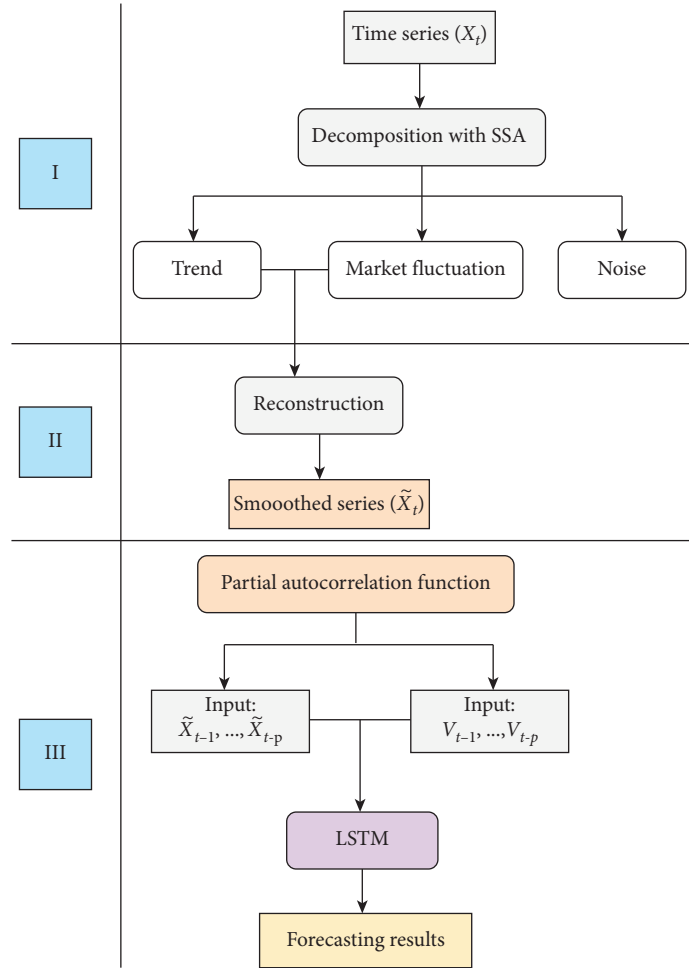


FIGURE 5: Hybrid SSA-LSTM model processing process.

lag of \tilde{x}_t and V_t is determined. Then, the final prediction result is obtained.

SSA is used as a pretreatment method to extract effective information of overall tendency and market volatility from primary sequential data.

This paper chooses $m = 10$, that is, the original series is decomposed into ten layers. According to the singular values shown in Table 1, it is found that, for this financial high-frequency time series, the first layer already contains more than 99.99% of the sequence information. Therefore, the first layer is selected as the reconstruction of the sequence, and the rest are regarded as noise.

2.5. Training Method and Optimizer Selection. The goal of this article is to compare the prediction effect of the closing price of DJIA, so mean square error (MSE) is chosen as the loss function. As for the optimizer, since the Adam algorithm has advantages over other adaptive learning rate algorithms in convergence speed and learning effect, this article uses the Adam optimizer (Adaptive Moment Estimation) for optimization training. Epochs are set to be 10. This article is based on the *Python* language environment and uses TensorFlow as the deep

TABLE 1: Singular values of singular spectrum analysis.

i	λ_i
1	0.999991285112503
2	$5.18930589813891e-06$
3	$1.42616024536684e-06$
4	$6.95916781349973e-07$
5	$4.21779000682764e-07$
6	$2.87333215462852e-07$
7	$2.15658198719760e-07$
8	$1.77801416057950e-07$
9	$1.55528979155002e-07$
10	$1.45403762506482e-07$

learning framework for training, prediction, and comparison.

3. Data and Results

3.1. Data. In order to research the feasibility and effectiveness of the denoising methods to forecast the actual financial sequential data, this section compares the prediction results of the RNN and LSTM models with a dropout layer and the LSTM model using data denoising methods

with the prediction results of the original LSTM model. The data sample has selected the latest data available in the past years. That is, the time interval is from 00:00 on January 1, 2020, to 23:55 on December 31, 2020, with a time interval of 5 minutes and a total of 19,666 pieces of data. And descriptive statistics of the data are shown in Table 2. The data come from the Wind database.

3.2. Test Set Prediction Effect Evaluation Index. To verify the validity of the model, the prediction and verification have been carried out from the three dimensions of short term (1 hour), medium term (3 hours), and long term (6 hours). Root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) are used as the prediction accuracy indexes to evaluate the predictive effect of the test set. The smaller the values of the above three indicators are, the higher the prediction accuracy is. The prediction stability is evaluated by the standard deviation of absolute percentage error (SDAPE). The lower the SDAPE value, the higher the reliability of prediction.

$$\begin{aligned}
 \text{RMSE} &= \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2}, \\
 \text{MAE} &= \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|, \\
 \text{MAPE} &= \frac{1}{N} \sum_{i=1}^N \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100, \\
 \text{SDAPE} &= \sqrt{\frac{1}{N} \sum_{i=1}^N \left(\frac{y_i - \hat{y}_i}{y_i} - \text{MAPE} \right)^2}.
 \end{aligned} \tag{9}$$

Here, y_i is the actual value and \hat{y}_i is prediction derived from the forecast model. N is the number of predictions.

3.3. Comparative Analysis of Short-Term Forecasting Effects. For the short-term prediction effect, according to the results shown in Table 3, in terms of prediction accuracy, RNN with a dropout layer is superior to the LSTM model, and RMSE, MAE, and MAPE are decreased by 11.63%, 1.96%, and 1.89%, respectively. Besides, the dropout layer [21] can optimize the prediction effect of the original LSTM model, and RMSE, MAE, and MAPE are decreased by 34.81%, 37.95%, and 37.95%, respectively. SSA-LSTM can optimize the prediction effect of the original LSTM model, and RMSE, MAE, and MAPE are reduced by 70.11%, 65.73%, and 65.72%, respectively. WT-LSTM can optimize the prediction effect of the original LSTM model, reducing RMSE, MAE, and MAPE by 79.55%, 76.91%, and 76.91%, respectively. In terms of prediction stability, RNN with a dropout layer is also superior to the LSTM model and the SDAPE is reduced by 28.24%. Besides, the dropout layer can improve the prediction stability of the original LSTM model and reduce the SDAPE by 30.41%. SSA-LSTM can improve the

TABLE 2: Descriptive statistics of DJIA closing price every 5 minutes.

Obs	Mean	Std. dev	Variance	Skewness	Kurtosis
19666	26876	2513.763	6319004	-0.8633175	3.213287

TABLE 3: 1-hour DJIA closing price forecast results.

	RMSE	MAE	MAPE	SDAPE
LSTM	5.8516916	4.5195833	0.0001481	0.0001218
RNN-dropout	5.1708886	4.4310000	0.0001453	0.0000874
LSTM-dropout	3.8146496	2.8042500	0.0000919	0.0000848
SSA-LSTM	1.7488158	1.5490332	0.0000508	0.0000266
WT-LSTM	1.1966503	1.0434276	0.0000342	0.0000192

prediction stability of the original LSTM model and reduce the SDAPE by 78.14%. WT-LSTM can improve the prediction stability of the original LSTM model and reduce the SDAPE by 84.23%. Through the analysis of short-term prediction, it is found that the generalization effect of filtering to prevent overfitting is better than the effect of the dropout layer to improve the accuracy. At the same time, wavelet transform has a good filtering effect. In terms of prediction accuracy, WT-LSTM can improve the prediction effect of SSA-LSTM model, and RMSE, MAE, and MAPE are reduced by 31.57%, 32.64%, and 32.64%, respectively. In terms of predictive stability, WT-LSTM can improve the predictive stability of the SSA-LSTM model, and SDAPE can be reduced by 27.85%. The prediction results of the four methods are shown in Figures 6–10.

3.4. Comparative Analysis of Medium-Term Forecasting Effects. For the medium-term prediction effect, according to the results shown in Table 4 in terms of prediction accuracy, RNN with a dropout layer is superior to the LSTM model, and RMSE, MAE, and MAPE are decreased by 10.64%, 4.54%, and 4.53%, respectively. Besides, the dropout layer can optimize the prediction effect of the original LSTM model, and RMSE, MAE, and MAPE are decreased by 34.54%, 36.15%, and 36.14%, respectively. SSA-LSTM can optimize the prediction effect of the original LSTM model, and RMSE, MAE, and MAPE are reduced by 77.67%, 76.49%, and 76.49%, respectively. WT-LSTM can optimize the prediction effect of the original LSTM model, reducing RMSE, MAE, and MAPE by 75.60%, 73.29%, and 73.28%, respectively. In terms of prediction stability, RNN with a dropout layer is also superior to the LSTM model and the SDAPE is reduced by 21.23%. Besides, the dropout layer can improve the prediction stability of the original LSTM model and reduce the SDAPE by 32.03%. SSA-LSTM can improve the prediction stability of the original LSTM model and reduce the SDAPE by 79.20%. WT-LSTM can improve the prediction stability of the original LSTM model and reduce the SDAPE by 79.79%. Through the analysis of medium-term prediction, we can see that the generalization effect of filtering to prevent overfitting is better than the effect of the dropout layer to improve the accuracy. At the same time, singular spectrum analysis also has a good effect on filtering.

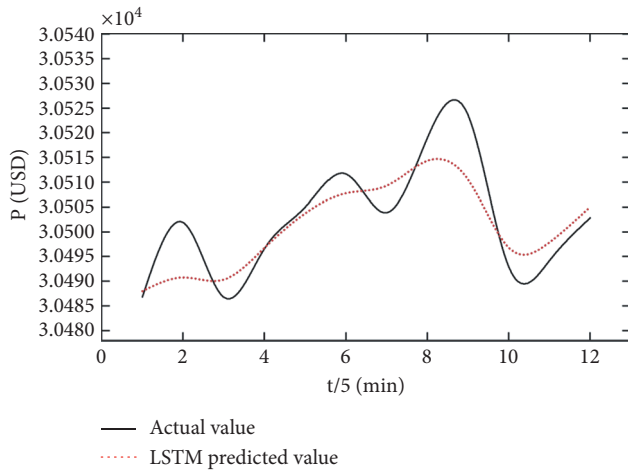


FIGURE 6: Short-term LSTM prediction.

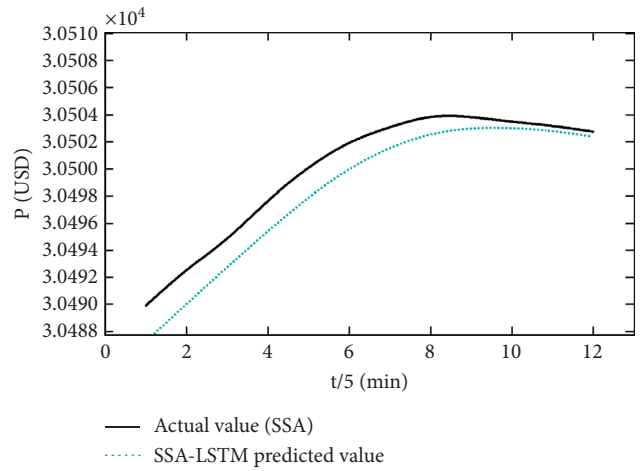


FIGURE 9: Short-term SSA-LSTM prediction.

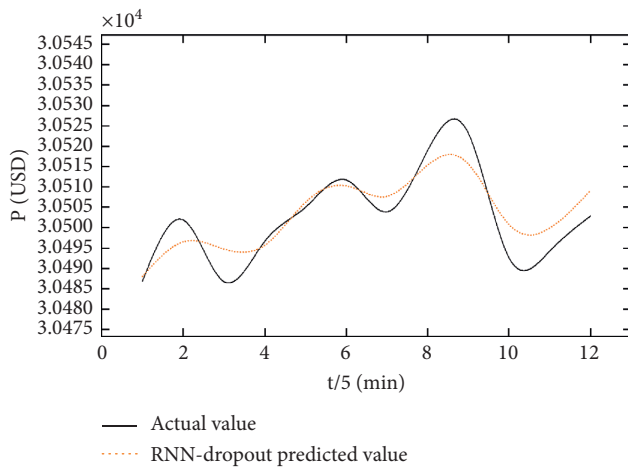


FIGURE 7: Short-term RNN-dropout prediction.

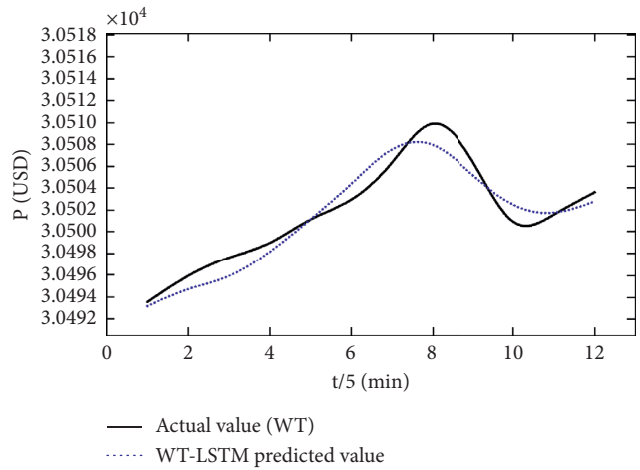


FIGURE 10: Short-term WT-LSTM prediction.

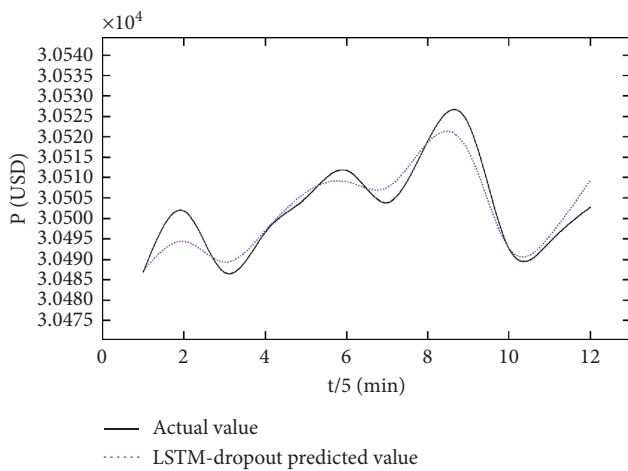


FIGURE 8: Short-term LSTM-dropout prediction.

TABLE 4: 3-hour DJIA closing price forecast results.

	RMSE	MAE	MAPE	SDAPE
LSTM	5.2447743	4.1066389	0.0001347	0.0001069
RNN-dropout	4.6867001	3.9202222	0.0001286	0.0000842
LSTM-dropout	3.4334542	2.6221944	0.0000860	0.0000727
SSA-LSTM	1.1713269	0.9653596	0.0000317	0.0000222
WT-LSTM	1.2796409	1.0970283	0.0000360	0.0000216

respectively. Concerning predictive stability, WT-LSTM can improve the predictive stability of the SSA-LSTM model and reduce SDAPE by 2.84%. The predicted results of the four methods are shown in Figures 11–15.

In terms of prediction accuracy, SSA-LSTM can improve the prediction effect of the WT-LSTM model and reduce RMSE, MAE, and MAPE by 8.46%, 12.00%, and 12.00%,

3.5. Comparative Analysis of Long-Term Forecasting Effects. For long-term forecasts, according to the results shown in Table 5, in terms of prediction accuracy, RNN with a dropout layer is superior to the LSTM model, and RMSE,

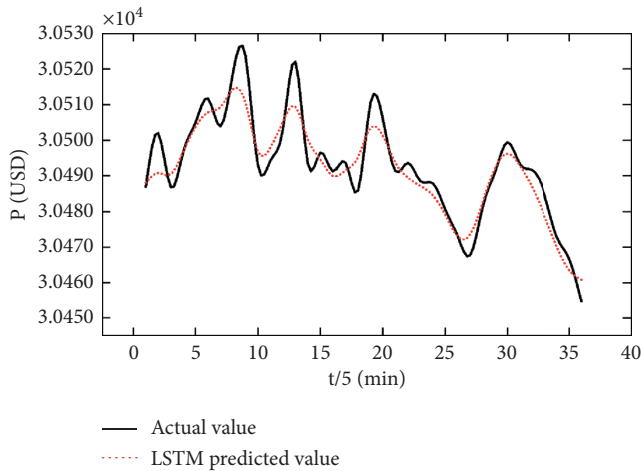


FIGURE 11: Medium-term LSTM prediction.

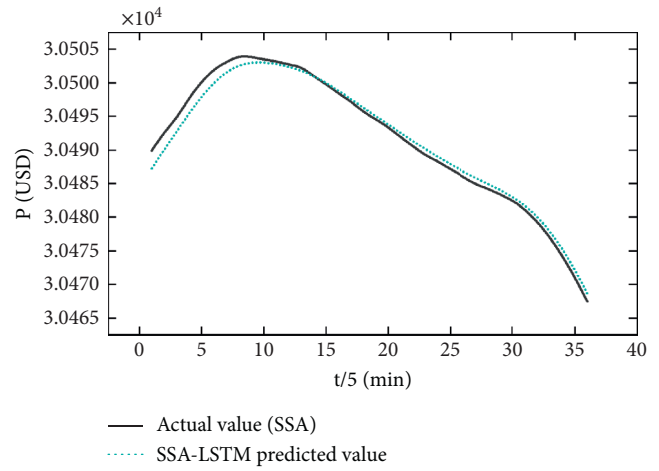


FIGURE 14: Medium-term SSA-LSTM prediction.

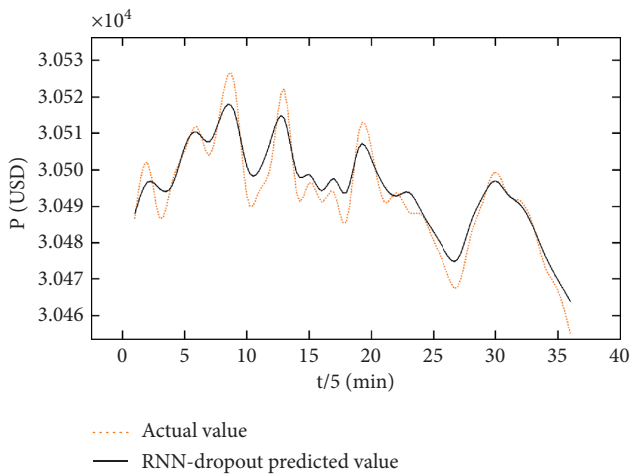


FIGURE 12: Medium-term RNN-dropout prediction.

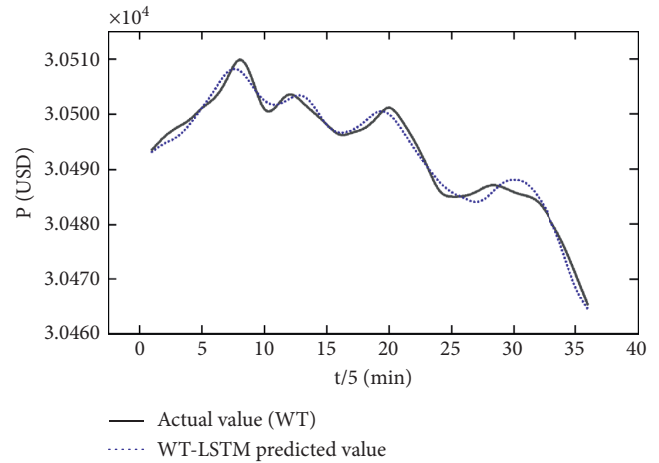


FIGURE 15: Medium-term SSA-LSTM prediction.

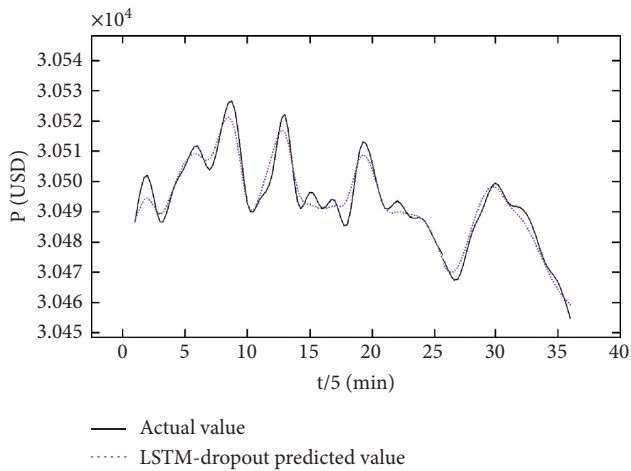


FIGURE 13: Medium-term LSTM-dropout prediction.

MAE, and MAPE are decreased by 13.02%, 6.03%, and 5.99%, respectively. Besides, the dropout layer can optimize the prediction effect of the original LSTM model, and RMSE,

MAE, and MAPE are decreased by 26.99%, 29.56%, and 29.54%, respectively. SSA-LSTM can optimize the prediction effect of the original LSTM model, and RMSE, MAE, and MAPE are reduced by 80.94%, 78.28%, and 78.28%, respectively. WT-LSTM can optimize the prediction effect of the original LSTM model, reducing RMSE, MAE, and MAPE by 68.92%, 69.15%, and 69.14%, respectively. In terms of prediction stability, RNN with a dropout layer is also superior to the LSTM model and the SDAPE is reduced by 22.35%. Besides, the dropout layer can improve the prediction stability of the original LSTM model and reduce the SDAPE by 23.90%. SSA-LSTM can improve the prediction stability of the original LSTM model and reduce the SDAPE by 84.82%. WT-LSTM can improve the prediction stability of the original LSTM model and reduce the SDAPE by 68.57%. According to the analysis of long-term prediction, we can see that the generalization effect of filtering to prevent overfitting is better than the effect of the dropout layer to improve the accuracy. At the same time, singular spectrum

TABLE 5: 6-hour DJIA closing price forecast results.

	RMSE	MAE	MAPE	SDAPE
LSTM	6.1655946	4.5780000	0.0001503	0.0001356
RNN-dropout	5.3630017	4.3020694	0.0001413	0.0001053
LSTM-dropout	4.5014469	3.2249583	0.0001059	0.0001032
SSA-LSTM	1.1753464	0.9942363	0.0000326	0.0000206
WT-LSTM	1.9164739	1.4123594	0.0000464	0.0000426

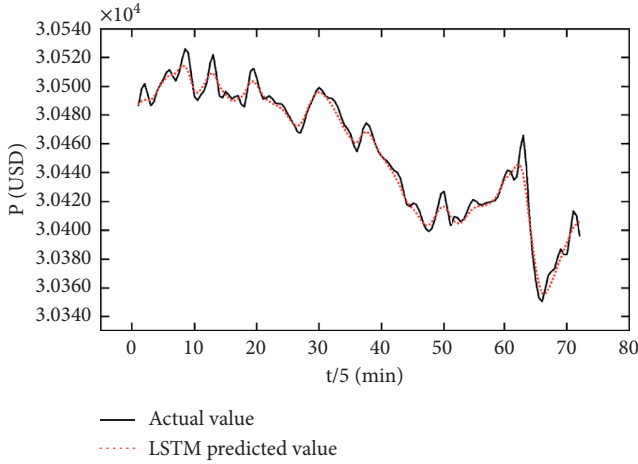


FIGURE 16: Long-term LSTM prediction.

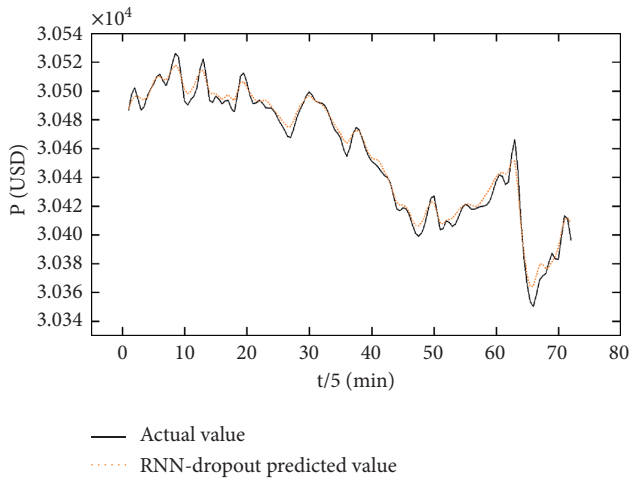


FIGURE 17: Long-term RNN-dropout prediction.

analysis also has a good effect on filtering. In terms of prediction accuracy, SSA-LSTM can improve the prediction effect of the WT-LSTM model and reduce RMSE, MAE, and MAPE by 38.67%, 29.60%, and 29.63%, respectively. In terms of predictive stability, SSA-LSTM can improve the predictive stability of WT-LSTM model and reduce SDAPE by 51.70%. The predicted results of the four methods are shown in Figures 16–20.

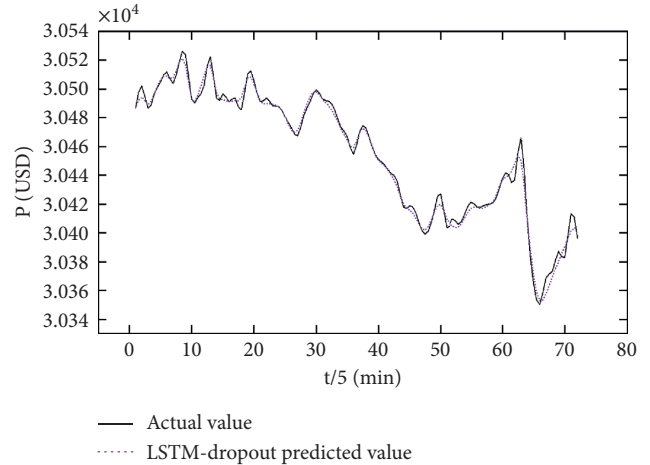


FIGURE 18: Long-term LSTM-dropout prediction.

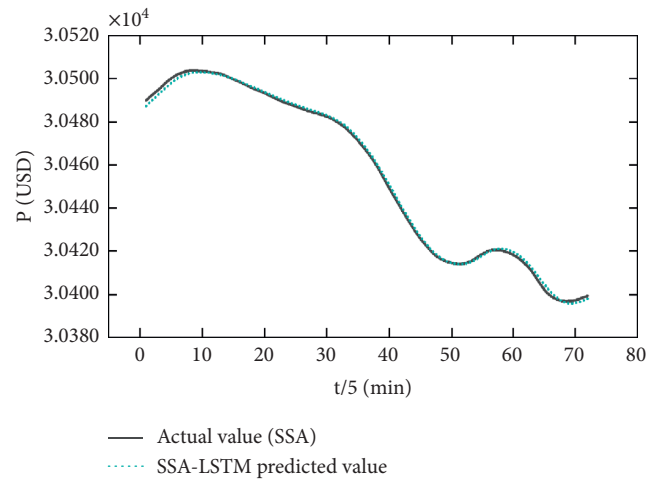


FIGURE 19: Long-term SSA-LSTM prediction.

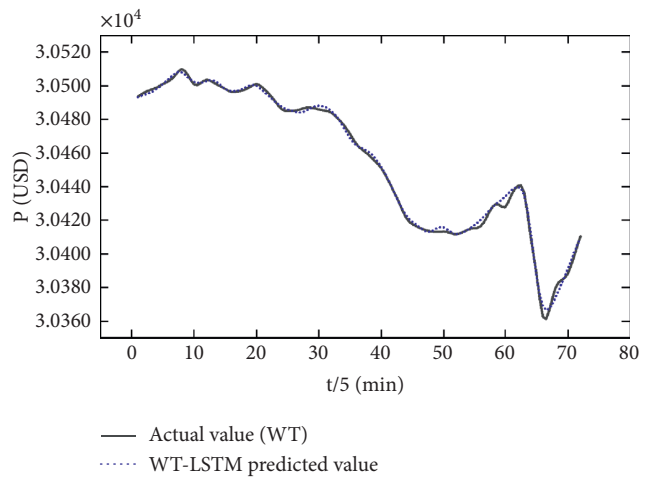


FIGURE 20: Long-term WT-LSTM prediction.

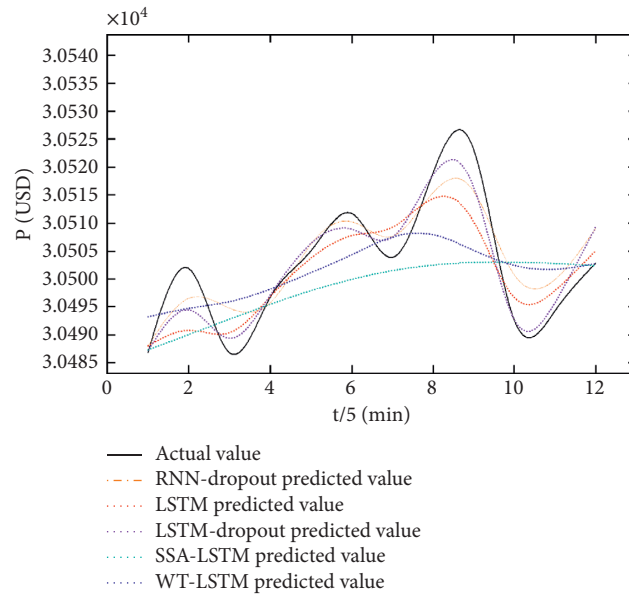


FIGURE 21: Short-term prediction results.

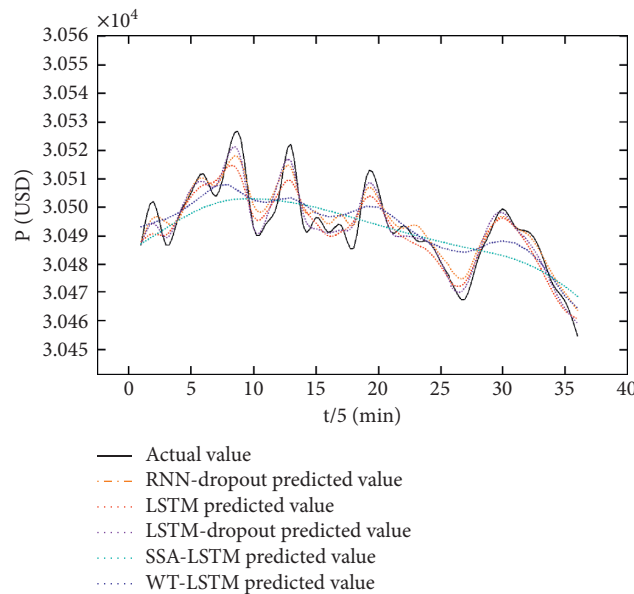


FIGURE 22: Medium-term prediction results.

In summary, both WT-LSTM and SSA-LSTM can significantly enhance the prediction ability of the original LSTM and raise the prediction accuracy and stability, especially the generalization ability, no matter in the short, medium, or long term. Also, in the short term and medium

term, the improvement effect of WT-LSTM is better than that of SSA-LSTM, while in the long term, the improvement effect of SSA-LSTM is better than that of WT-LSTM. A comprehensive comparison of the prediction results of the four methods is shown in Figures 21–23.

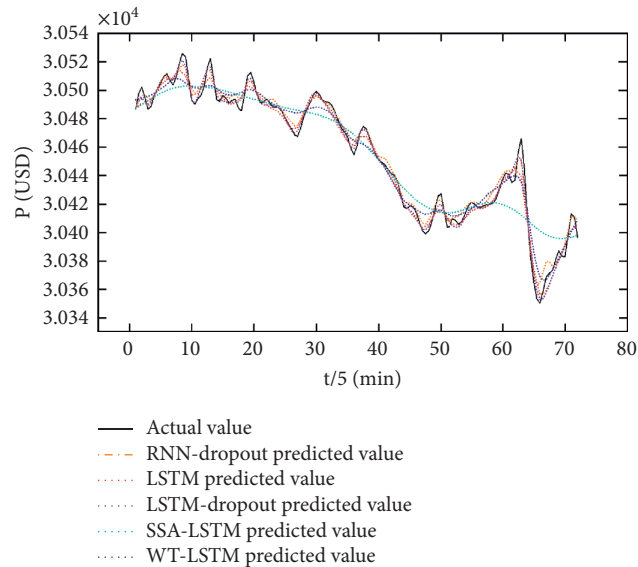


FIGURE 23: Long-term prediction results.

4. Conclusion and Discussions

This paper has discussed the theoretical basis of deep learning and the practical application of LSTM price prediction and has proposed the use of denoising methods to reduce noise on high-frequency financial time series to minimize the effect of random interference noise to raise the prediction generalization of the model for out-of-sample data. The results are also significant enough to prove the improvement of the LSTM predicting model with effective denoising methods, especially wavelet transform and singular spectrum analysis.

In light of the empirical results of the DJIA 5 minutes closing data, the following conclusions can be drawn: firstly, the use of wavelet transform and singular spectrum analysis to denoise data can significantly raise the ability to extend of LSTM neural network and WT's effect is better than SSA's effect in the short term and medium term, but worse than SSA filtering method in the long term. Secondly, with the extension of the time limit, the generalization ability of wavelet transform and reconstructed filter sequence in the prediction of LSTM neural network is weakening, while the generalization ability of singular spectrum analysis decomposition and reconstructed filter sequence in the prediction of LSTM neural network is increasing. But the prediction effect of wavelet transform and singular spectrum analysis reconstruction filter is still significant. Thirdly, the WT-LSTM neural network and SSA-LSTM neural network can converge quickly in a small amount of time and has a good prediction effect under the high-frequency data, which provides a new idea for financial risk management and monitoring under high-frequency trading. These findings can be widely used in the selection of methods for processing time series. To be specific, WT-LSTM is recommended to be chosen when processing relatively short-term time series, while SSA-LSTM is more efficient in processing long-term time series.

In view of the high tunability of neural networks, there are still many technical improvements in future research, such as adding more nonhomogeneous information as input to the neural network and optimizing the structure of the neural network itself. It is worth noting that applying the advantages of big data in the financial high-frequency time series to the investment field can enable investment opportunities to be discovered by investors in a timely manner and facilitate the development of intelligent investment in the financial market. In addition, it can also strengthen risk management, improve the efficiency of risk identification, and effectively maintain the stability of the financial market.

Data Availability

The DJIA data used to support the findings of this study were supplied by Wind under license and so cannot be made freely available. Requests for access to these data should be made to Wind, sales@wind.com.cn.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments





The authors gratefully acknowledge Southeast University for providing them with rich academic resources. At the same time, thanks are due to Dr. Wei Yin for the inspiration of this work and careful instruction and to Dr. Yawen Han for valuable suggestions. The authors are also grateful to the partners for their full cooperation. In addition, Qi Tang wants to thank Wenye Shi and Yuanyu Wang for their assistance and encouragement, as well as the concern and company from Ciyue Han for years.

References

- [1] X. Zheng and B. M. Chen, "Analysis of Dow Jones industrial average," *Stock Market Modeling and Forecasting*, vol. 442, pp. 43–51, 2013.
- [2] W. Xu, H. Peng, F. Tian, and X. Peng, "A hybrid modelling method for time series forecasting based on a linear regression model and deep learning," *Applied Intelligence*, vol. 49, no. 8, pp. 3002–3015, 2019.
- [3] N. K. Ahmed, A. F. Atiya, N. E. Gayar, and H. El-Shishiny, "An empirical comparison of machine learning models for time series forecasting," *Econometric Reviews*, vol. 29, no. 5-6, pp. 594–621, 2010.
- [4] G. Sun, J. Lin, C. Yang et al., "Stock price forecasting: an echo state network approach," *Computer Systems Science and Engineering*, vol. 36, no. 3, pp. 509–520, 2021.
- [5] J. S. Cao and J. H. Wang, "Stock price forecasting model based on modified convolution neural network and financial time series analysis," *International Journal of Communication Systems*, vol. 32, 2019.
- [6] N. Koç Ustali, N. Tosun, and Ö. Tosun, "Makine öğrenmesi teknikleri ile hisse senedi fiyat tahmini," *Eskişehir Osmangazi Üniversitesi İktisadi Ve İdari Bilimler Dergisi*, vol. 16, no. 1, pp. 1–16, 2021.
- [7] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [8] Q. Jiang, C. Tang, C. Chen, X. Wang, and Q. Huang, "Stock price forecast based on LSTM neural network," in *Twelfth International Conference on Management Science and Engineering Management*, pp. 393–408, Springer, Berlin, Germany, 2018.
- [9] N. E. Huang, Z. Shen, S. R. Long et al., "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society of London. Series A: Mathematical, Physical And Engineering Sciences*, vol. 454, no. 1971, pp. 903–995, 1998.
- [10] H. Hassani, "Singular spectrum analysis: methodology and comparison," *Journal of Data Science*, vol. 5, pp. 239–257, 2007.
- [11] C.-S. Lin, S.-H. Chiu, and T.-Y. Lin, "Empirical mode decomposition-based least squares support vector regression for foreign exchange rate forecasting," *Economic Modelling*, vol. 29, no. 6, pp. 2583–2590, 2012.
- [12] L. H. Tang, Y. L. Bai, J. Yang et al., "A hybrid prediction method based on empirical mode decomposition and multiple model fusion for chaotic time series," *Chaos Solitons & Fractals*, vol. 141, 2020.
- [13] J. Wang and X. Li, "A combined neural network model for commodity price forecasting with SSA," *Soft Computing*, vol. 22, no. 16, pp. 5323–5333, 2018.
- [14] F. Wen, J. Xiao, Z. He, and X. Gong, "Stock price prediction based on SSA and SVM," *Computer Science*, vol. 31, pp. 625–631, 2014.
- [15] L. Yu, S. Wang, and K. K. Lai, "Forecasting crude oil price with an EMD-based neural network ensemble learning paradigm," *Energy Economics*, vol. 30, no. 5, pp. 2623–2635, 2008.
- [16] N. Jia and C. J. Zheng, "Short-term forecasting model of agricultural product price index based on LSTM-DA neural network," *Computer Science*, vol. 46, pp. 62–65, 2019.
- [17] H. T. Jung, H. H. Fen, and Y. W. Chang, "Forecasting stock markets using wavelet transforms and recurrent neural networks: an integrated system based on artificial bee colony algorithm," *Applied Soft Computing*, vol. 11, pp. 2510–2525, 2011.
- [18] Q. Tang, T. M. Fan, R. C. Shi, J. Y. Huang, and Y. Y. Ma, "Prediction of financial time series using lstm and data denoising methods," <http://arxiv.org/abs/2103.03505>.
- [19] Z. Y. Pang and L. Liu, "Can agricultural product price fluctuation be stabilized by future market: empirical study based on discrete wavelet transform and GARCH model," *Financial Research*, vol. 11, pp. 126–139, 2013.
- [20] X. L. Liu and X. P. Wang, "High frequency prediction research based on wavelet analysis of stock index futures," *Systems Engineering Theory and Practice*, vol. 35, pp. 1425–1432, 2015.
- [21] N. Srivastava, G. E. Hinton, A. Krizhevsky et al., "Dropout : a simple way to prevent networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

Research Article

Prefabricated Concrete Component Geometry Deviation Statistical Analysis

Xiaoyong Luo ^{1,2}, Hao Long ^{1,2}, Shuang Dong ¹ and Jingyi Wu ³

¹Department of Civil Engineering, Central South University, Changsha 41000, China

²Hunan Prefabricated Construction Engineering Technology Research Centre, Changsha 41000, China

³Hatchip Co., Shenzhen 518000, China

Correspondence should be addressed to Hao Long; longhao@csu.edu.cn

Received 22 March 2021; Revised 26 April 2021; Accepted 5 May 2021; Published 19 May 2021

Academic Editor: Jie Chen

Copyright © 2021 Xiaoyong Luo et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The research objects of this paper were the prefabricated concrete components produced by four enterprises in China, and the dimension deviation data of more than 1400 prefabricated concrete components are measured with high-precision 3D photogrammetry technology. The nonparametric Kruskal–Wallis test was carried out for the size deviation of the same type of components produced by different enterprises. The distribution characteristics of geometric parameters of typical components of prefabricated structures in China, such as beams, columns, wall boards, and composite slabs, were analyzed by using the probability statistical method. The Kolmogorov–Smirnov goodness-of-fit method was used to test the cumulative distribution function of dimension deviation, and the size distribution of fabricated components was studied. The results showed that there was no significant difference in the size deviation of the same-type component produced by different enterprises, and the range of geometric parameter uncertainty random variables was small, which was between 0.99 and 1.02. Also, the fluctuation was small, the coefficient of variation was below 0.0093, and the variability of component size deviation was small. The transverse dimension of the component shows a positive deviation, the vertical dimension of component shows a negative deviation, and the dimension deviation of prefabricated concrete components follows the normal distribution.

1. Introduction

Prefabricated building is one of the important directions of building structure development, which is conducive to the development of building industrialization. As a crucial part of the prefabricated concrete (PC) structure, the quality of prefabricated components plays an important role in the overall structural performance [1]. In terms of the splicing structure of prefabricated components, if the size of prefabricated components is too large, it would be difficult to assemble and if the size is too small, the joints between components would be too large, causing water seepage problems. From the aspect of stress performance, when analyzing the reliability of the structure according to the probability limit state design method, the failure probability of the structural part or its corresponding reliability index is used as the measurement, and the geometrical dimension

variable is one of the aspects that affect the resistance of the structural part. Moreover, contemporary buildings have become lighter and more vulnerable to building movement and subsequent geometric changes [2]. The lack of uniformity of accuracy between factory-made and in-situ components and the higher level of building movements in the contemporary building are two major factors that affect the dimensional and geometric accuracy of buildings [3]. The conversion of a good design into a good product (e.g., a building) is a matter of keeping dimensional and geometric variations within tolerances that are predetermined at the design stage [4]. The acceptability of a product depends on whether its variations in size and geometry fall within set limits; thus, the bridge between design and production is tolerance. In other words, tolerances interlink design with construction because without specifying the tolerances, it is not clear whether components and subassemblies (i.e.,

connections of two or more components) meet the design intent regarding the accuracy of the final product [5].

At present, there are many quality control standards for fabricated components, such as German DIN18203 and DIN18202 standards, Japanese JASS10, American MNL-116 and ACI regulations, and Chinese standards GB/T51231-2016 and GB/T 51232-2016 [6–10]. Each standard specifies the dimensional deviation of prefabricated components. These standards usually determine the allowable value of dimensional deviation based on traditional experiences [11], and the deviation allowable value for prefabricated components is strictly controlled and appropriately modified based on the relevant specifications of cast-in-place concrete. Scholars do not pay attention to the prefabricated component manufacturers' production capacity and have not systematically investigated the dimensional deviation of prefabricated components [12]. To determine the reliability of the concrete structure design, in the 1970s and 1980s [13], a survey of the size deviation of cast-in-place concrete structures was conducted in China [13]. Liu et al. [11] conducted the statistical investigation on the size deviation of cast-in-place concrete structures in Beijing and suggested adjusting the size deviation acceptance index, while Mao and Shi [14] conducted the statistical investigation and analysis on the section size and axis position of cast-in-place concrete structures, and they believed that the deviation of concrete section size obeys the normal distribution, yet the axis position deviation obeys the logarithmic distribution. Nowadays, there are many types of research on the size deviation of cast-in-situ structures, but there are quite few statistical studies on the size deviation of prefabricated concrete components. Therefore, it is of great significance to systematically investigate the size deviation of fabricated components.

In this paper, the dimensions of fabricated concrete components of different companies in China are measured by 3D photogrammetry technology, obtaining a database of wall board, composite slab, beam, and column size deviation. The dimensional deviation distribution characteristics of fabricated components are analyzed.

2. Data Acquisition

2.1. Overview of Sample Acquisition. The focus is mainly on the component forms existing in prefabricated building structures: beams, columns, wall panels, and composite slabs. In terms of survey objects, components produced by prefabricated component factories in different regions of China are mainly selected. Random sampling is used for each type of component. Detailed information about the number of survey samples is shown in Table 1.

2.2. Measurement Method. Prefabricated building production and installation have higher precision requirements than traditional cast-in-situ concrete structures. The traditional measurement method for prefabricated concrete buildings is mainly using steel rulers, and there are problems with the use of steel rulers in measuring the dimensions of

prefabricated concrete components, such as insufficient accuracy, large errors, and difficulty in measuring the edge corners of the component dimensions. The HL-3DP is a medium-range photogrammetry scanner type. It was designed for medium-space applications such as building, civil and survey, and forensics. Three-dimensional photogrammetry known as HL-3DP was applied for prefabricated concrete, and the scanned data were managed with their associated software HOLON3DP. HL-3DP is a high-accuracy and medium-range photogrammetry system. It is associated with a high-resolution camera. Table 2 shows the specification of HL-3DP. Therefore, in this article, HL-3DP three-dimensional photogrammetry was adopted to collect sample data. Figure 1 shows the on-site measurement of prefabricated components. The relative error of this method is less than 5%, and the absolute error is less than 0.01 mm. The measurement of the size of the prefabricated component, thus, resulted in high precision in terms of both experimental error and measurement accuracy.

2.3. Accuracy Assessment. Accuracy assessment was also performed in this study. The findings show that the 3D photogrammetry of HL-3DP can provide high accuracy of the building structure. To get a systematic accuracy assessment, the evaluation of the estimate accuracy was carried out using equation (1), where x represents each value in the dataset, \bar{x} represents the mean of all values in the dataset, and n represents the number of value in the dataset.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (x - \bar{x})^2}{n}}. \quad (1)$$

The root-mean-square error (RMSE) was used to measure the differences between values observed. According to Mao and Shi [14], RMSE is commonly used in the research field to describe the accuracy of features and it is acceptable to measure the error and estimate the quality of features. The lower the RMSE value, the better the accuracy.

Table 3 shows the dimensional measurement difference between the measuring steel tape and photogrammetry system. Meanwhile, the analysis for this study included the comparison of measurement between the photogrammetry system and conventional method (measuring steel tape) with the design value. The comparison is performed and recorded in Table 4. The RMSE value of the photogrammetry system data was 0.60 mm while that of the measuring steel tape method was 0.94 mm. As the result, the tolerance of the photogrammetry system model is well within the 0.01-millimeter level. The photogrammetry system model has a lower tolerance compared to the measuring steel tap. This proved that the photogrammetry system method was accurate.

3. Results of Sample Collection

In the process of statistical analysis of the geometric parameters of the components, because the design dimensions of the components are not the same, to analyze and compare the variation of the dimensions described by the uncertainty

TABLE 1: Sample number of prefabricated concrete members in different enterprises.

Survey object	Sample number of prefabricated parts			
	Beam components	Column components	Wall panel components	Composite slab components
Enterprise 1	62	66	102	87
Enterprise 2	54	129	84	96
Enterprise 3	76	94	64	111
Enterprise 4	103	85	121	107

TABLE 2: Specification of 3D photogrammetric HL-3DP.

HL-3DP 3D photogrammetric instrument	
Range measurement principle scanner control	Measurement of noncontact optical tricoordinate
Range	Up to 20 m (minimum range 0.01 m)
Field of view	360°horizontal, 360°vertical
Signal image	1.8 MB, 4Mpixels (1920 × 1920 pixel)
Ranging error	±0.01 mm



(a)



(b)

FIGURE 1: 3D photogrammetric component size. (a) Composite slab 3D measurement. (b) 3D measurement of prefabricated columns.

of the geometric parameters with the random variable K_A [15], which can be expressed as shown in equation (2), where a is the actual value of the geometric parameter of the component and a_K is the standard value of the geometric parameter of the component (usually the design value),

$$K_A = \frac{a}{a_K}. \tag{2}$$

When analyzing the variability of the geometric dimensions of components, variables such as length, width, and height are regarded as independent random variables.

Taking the same type of component produced by the same enterprise as a matrix, its statistical parameters are analyzed with a subsample of capacity $n > 50$. Integrating the statistical analysis results of various enterprises, the weighted average of the same type of component size variation is used to reflect the variation level of geometric dimensions. This paper is based on the measured data, and the number of subsamples participating in the statistical analysis is the weight (the capacity is n) [16]. Statistical parameters of random variables with uncertain geometric dimensions of various components were obtained as follows:

TABLE 3: Result of dimensional measurement.

Prefabricated	Dimension	Design value (mm)	Measuring steel tape (mm)	Photogrammetry system (mm)
Prefabricated beam components	Length	2830	2833	2831.86
	Width	400	402	400.82
	Height	410	413	412.03
Prefabricated column components	Length	2535	2539	2537.08
	Width	500	503	502.07
	Height	520	523	520.67
Prefabricated wall panel components	Height	6630	6633	6629.84
	Width	2600	2598	2599.24
	Thickness	240	238	240.65
Prefabricated composite slab*	Length	4350	4355	4346.06
	Width	3930	3933	3926.95

*Since the thickness direction of the prefabricated composite slab is cast-in-place concrete, it was not counted.

TABLE 4: Comparison of measurement between the measuring steel tape and photogrammetry system.

Prefabricated	Measuring steel tape (mm)	x^2	Photogrammetry system	x^2
Prefabricated	-3	9	-1.86	3.46
Beam components	-2	4	-0.82	0.67
	-3	9	-2.03	4.12
Prefabricated	-4	16	-2.08	4.33
Column	-3	9	-2.07	4.28
Components	-3	9	-0.67	0.45
Prefabricated wall	-3	9	0.16	0.02
Panel components	2	4	0.76	0.58
	2	4	-0.65	0.42
Prefabricated	-5	25	3.94	15.52
Composite slab	-3	9	3.05	9.30
Sum	Σ	107	Σ	43.15
RMSE	$\sqrt{\Sigma}/11$	0.94	$\sqrt{\Sigma}/11$	0.60

$$\mu_{K_A} = \frac{\sum_{i=1}^n n_i \mu_{K_{Ai}}}{\sum_{i=1}^n n_i}, \quad (3)$$

$$\delta_{K_A} = \frac{\sum_{i=1}^n n_i \delta_{K_{Ai}}}{\sum_{i=1}^n n_i}, \quad (4)$$

where n_i is the subsample capacity of the geometric characteristic value of a certain construction unit, μ_{K_A} is the average value under capacity n_i , and δ_{K_A} is the coefficient of variation under n_i capacity. According to equations (2) and (3), the statistical parameters of the random variable of the geometric parameters of the component can be obtained.

3.1. Analysis of Acquisition Results. Taking prefabricated concrete components produced by different companies as the research object, analyzing the difference in geometric uncertainty of component size in different enterprises, and calculating the average production component K_A maximum, minimum, upper, and lower quartiles of each enterprise production, Figures 2–5 give a comparison of box plots. The figures show the different production enterprise production ranges of similar artifacts are

similar, both in the range of 0.99 to 1.02. The abnormal part of beam members accounted for 3.4%, the abnormal points of column members accounted for 0.8%, the abnormal percentage of wall slabs accounts for 3.4%, and the abnormal percentage of composite slab slabs accounts for 1.2%. There are fewer abnormal points, which indicates that the variability of component size deviation in the actual production process is small. The volume of the box body represents the dense and discrete degree of data distribution in the group. The length and width of beams, columns, wall panels, slabs, and width values K_A of the box body are relatively short, and the fluctuation is small.

Table 5 shows the comparison results of the K_A value among similar components of different manufacturers. The horizontal size of the components produced by different enterprises (wall panel height, wall panel width, composite slab length, and composite slab width) dimension uncertainty is less than 1, which is biased towards negative deviation, and longitudinal member dimensions are greater than 1, which are biased toward positive deviation. The size deviation of similar PC components produced by different companies is not large, which proves the reliability of the measured data. Manufacturers have less influence on the size of prefabricated PC

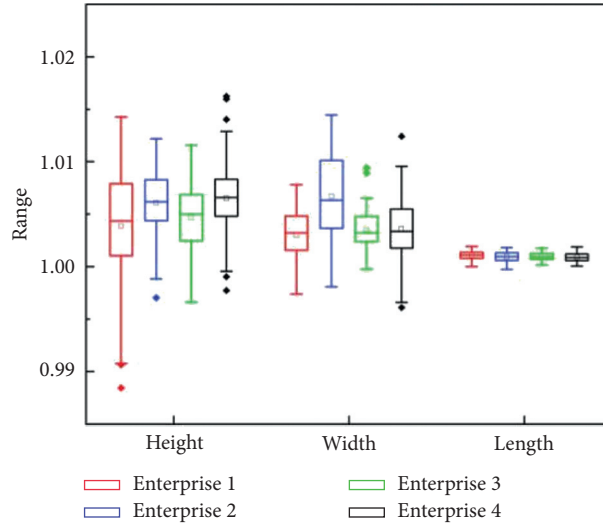


FIGURE 2: Box plot of geometric uncertainty of the beam member.

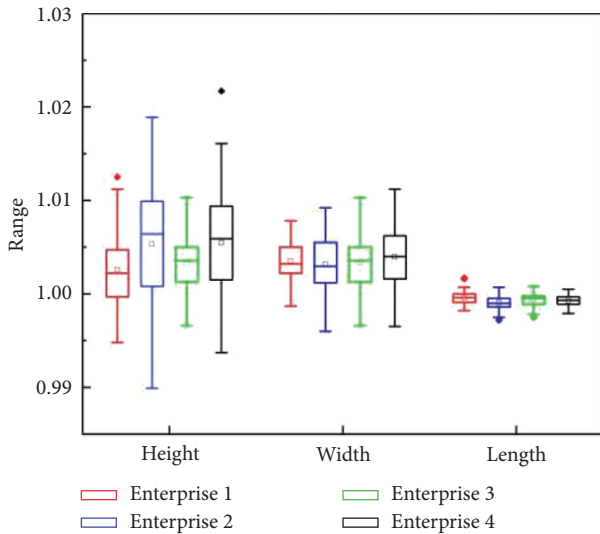


FIGURE 3: Box plot of geometric uncertainty of the column member.

components. The coefficients of variation K_A are below 0.0093, and the variation is small, indicating that similar components produced by different companies have the same size characteristics.

3.2. Correlation Test of Dimensional Deviation. With the significant level $\alpha = 0.05$ as the difference, a single factor analysis of variance on the dimensions of the components produced by the four companies is performed, the differences between the components produced by each factory are discussed, and it is evaluated whether the dimensional deviations of the components can be traced to a certain extent so that the measurement results of different component factories are comparable.

The Kruskal–Wallis test [17] infers whether the sample comes from the population median or the distribution pattern is significantly different. This method neither requires the data to

obey the normal distribution nor does it require the homogeneity of variance. The Kruskal–Wallis test is a nonparametric test method to test whether multiple population distributions are the same. Firstly, samples in multiple groups are mixed and sorted in ascending order to find the rank of each variable; if the values show no big difference, it can be considered that the distributions of multiple populations are not significantly different; on the contrary, if the rank means of each group are significantly different, the data of multiple groups cannot be mixed, and if the distributions of multiple samples are significantly different, the statistic constructed by the Kruskal–Wallis test is the K–W statistic, which is $p\bar{R}_i, \bar{R}$, shown as follows:

$$p = \frac{12}{N(N+1)} \sum_{i=1}^k n_i (\bar{R}_i - \bar{R})^2,$$

$$\bar{R}_i = \frac{R_i}{n_i}, \tag{5}$$

$$\bar{R} = \frac{N+1}{2},$$

where k is the number of sample groups, N is the total sample size, R_i is the sum of ranks of the group I , \bar{R}_i is the average rank of the group I , and \bar{R} is the total average rank.

Assuming that the distribution of the size deviation of similar components is the same in the company category, if $p < \alpha$, the null hypothesis is rejected, and the size deviation distributions of similar components produced by different companies have significant differences; on the contrary, if $p > \alpha$, then accept the null hypothesis, which means that the size deviation distribution of similar components produced by different companies has no significant difference. The nonparametric Kruskal–Wallis test was performed by SPSS19.0 software, and the analysis results are shown in Table 6. The significance value of each component size is greater than the significance level of 0.05, accepting the former hypothesis, so there is no significant difference in the size deviation of similar components produced by different

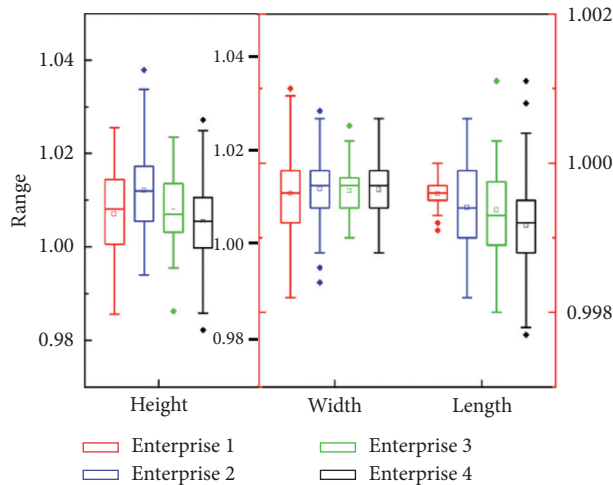


FIGURE 4: Box plot of wall panel geometric uncertainty.

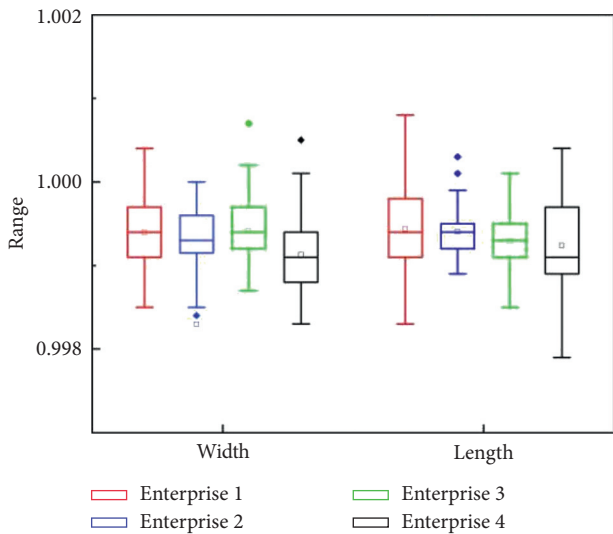


FIGURE 5: Box plot of geometric uncertainty of slab.

companies. The size deviation values of the same types of components produced by different companies can be used as subsamples to analyze the distribution characteristics of the size deviations of fabricated components in China.

4. Frequency Distribution of Size Deviation

Since there is no significant difference in the size deviation of the same type components, the size deviations of the same-type components produced by different companies are integrated as a sample of the size deviation, and the distribution law of size deviation is analyzed using histograms and box plots.

4.1. Dimensional Deviation Distribution of Prefabricated Beam Components. Calculating the difference between the measured data and the design value of the prefabricated beam, the sample size is 295, which is divided into 20 groups. The histogram is drawn with the deviation value as the abscissa and the measured frequency as the ordinate, as shown in Figure 6.

As seen in Figure 6, the dimensional deviation distribution histogram for the beam decreases from the center to the sides, and it can be seen from Figure 6 that the dimensional deviation distribution histogram of the beam is a normal distribution.

The length deviation of beam components mainly concentrated in the range of 0–6 mm, The most frequent deviation is 2–4 mm, the mean is 3.02 mm, and the variance is 1.91. The width deviation range is mainly concentrated in the range of 0–3 mm, the mean value is 1.57 mm, and the variance is 1.56, The height deviation is mainly concentrated in the range of 0–5 mm, the mean value is 2.44 mm, and the variance is 3.27.

The maximum, minimum, upper quartile, lower quartile, and median of the size deviations are compared using box plots, as shown in Figure 7. The characteristic values of the three-dimensional dimensional deviation of prefabricated beam components are all significantly different, and the dimensional deviations are basically positive deviations. There is no abnormal value in the length dimension deviation, the median is located in the center of the upper and lower quartiles, and the box chart is symmetrical about the median line, indicating that the length dimension deviation is symmetrically distributed. Both height and width deviations have abnormal values. The height deviation abnormal values are concentrated on the smaller value side and the median tends to the upper quartile, indicating that the height size deviation distribution is right-skewed distribution. The width size deviation abnormal values are concentrated on the larger value side and the median tends to the lower quartile, indicating that the width deviation distribution is left-skewed distribution.

4.2. Dimensional Deviation Distribution of Prefabricated Column Components. Calculating the difference between the measured data and the design value of the prefabricated column, the sample size is 374, which is divided into 20 groups. The histogram is drawn with the deviation value as the abscissa and the measured frequency as the ordinate, as shown in Figure 8. As seen in Figure 8, the dimensional deviation distribution histogram for the column decreases from the center to the sides, and it can be seen from Figure 8 that the dimensional deviation distribution histogram of the column is a normal distribution.

The length deviation of the column components mainly concentrated in the range of –5–1 mm, the most frequent deviation is –3–1 mm, the mean value is –2.05 mm, and the variance is 3.74; the width deviation mainly concentrated in the range of –1–3.5 mm, the most frequent deviation is 0.7–1.5 mm, the mean value is 1.34 mm, and the variance is 1.21; the height deviation is mainly concentrated in the range of –2–5 mm, the most frequent deviation is 1.5–2.5 mm, the average is 1.88 mm, and the variance is 4.34.

The maximum, minimum, upper quartile, lower quartile, and median of the size deviations are compared using box plots, as shown in Figure 9. The characteristic values of the three-dimensional dimensional deviation of prefabricated column components are all significantly different. There is no abnormal value for the width dimension deviation, the median deviation of width is located in the center of the upper and lower quartiles,

TABLE 5: Summary of sample statistical.

Prefabricated components	Project	μ_k/mm (δ_k)				Deviation range (mm)			
		Enterprise 1	Enterprise 2	Enterprise 3	Enterprise 4	Enterprise 1	Enterprise 2	Enterprise 3	Enterprise 4
Prefabricated beam member	Length	1.0011 (0.0004)	1.0009 (0.0005)	1.0010 (0.0004)	1.0065 (0.0034)	[0.04, 5.44]	[-0.77, 5.32]	[0.53, 5.52]	[0.16, 6.78]
	Width	1.0030 (0.0023)	1.0067 (0.0038)	1.0035 (0.0018)	1.0036 (0.0030)	[-1.05, 3.13]	[-0.81, 6.06]	[-0.09, 3.39]	[-1.48, 4.72]
	Height	1.0039 (0.0056)	1.0061 (0.0032)	1.0047 (0.0031)	1.0065 (0.0034)	[-4.74, 5.85]	[-1.25, 5.11]	[-1.56, 5.33]	[-1.09, 7.79]
Prefabricated column member	Length	0.9996 (0.0007)	0.9991 (0.0007)	0.9993 (0.0007)	0.9993 (0.0006)	[-4.53, 4.29]	[-8.06, 1.81]	[-6.54, 1.48]	[-6.16, 1.09]
	Width	1.0035 (0.0019)	1.0031 (0.0029)	1.0033 (0.0029)	1.0039 (0.0032)	[-0.25, 3.91]	[-1.44, 3.30]	[-1.30, 3.90]	[-1.25, 4.04]
	Height	1.0026 (0.0040)	1.0053 (0.0061)	1.0033 (0.0029)	1.0055 (0.0056)	[-1.51, 6.48]	[-3.65, 6.21]	[-2.28, 6.81]	[-2.19, 8.23]
Prefabricated wall panel components	Height	0.9996 (0.0002)	0.9994 (0.0005)	0.9994 (0.0006)	0.9992 (0.0006)	[-5.81, 0.24]	[-5.42, 2.16]	[-5.03, 1.15]	[-6.57, 3.29]
	Width	0.9996 (0.0005)	0.9997 (0.0004)	0.9996 (0.0003)	0.9997 (0.0004)	[-4.75, 2.29]	[-4.36, 2.05]	[-3.32, 1.77]	[-4.38, 2.04]
	Thickness	1.0071 (0.0093)	1.0121 (0.0089)	1.0076 (0.0070)	1.0054 (0.0084)	[-3.45, 5.88]	[-0.67, 5.98]	[-1.09, 4.48]	[-4.28, 5.98]
Prefabricated composite slab	Length	0.9994 (0.0006)	0.9994 (0.0003)	0.9993 (0.0003)	0.9992 (0.0005)	[-5.99, 3.10]	[-4.84, 1.12]	[-5.23, 0.23]	[-5.41, 0.92]
	Width	0.9994 (0.0004)	0.9993 (0.0003)	0.9994 (0.0004)	0.9991 (0.0004)	[-4.88, 1.22]	[-6.20, 0.09]	[-3.89, 2.13]	[-3.48, 1.10]

TABLE 6: Kruskal–Wallis inspection of component size deviation.

Prefabricated components	Project	N	R_i	\bar{R}_i	p	α^*
Prefabricated beam member	Length	295	584.07	146.02	0.131	0.05
	Width	295	611.65	152.91	0.177	
	Height	295	575.88	143.97	0.123	
Prefabricated column member	Length	368	765.93	191.48	0.401	
	Width	368	741.33	185.33	0.231	
	Height	368	727.64	181.91	0.090	
Prefabricated wall panel components	Height	404	818.47	204.61	0.061	
	Width	404	818.34	204.59	0.053	
	Thickness	404	821.64	205.41	0.077	
Prefabricated composite slab	Length	401	808.91	202.23	0.097	
	Width	401	796.64	199.16	0.081	

*Significance is progressive significance, and the significance level is 0.05.

and the square box chart is symmetrical about the median line, which means that the dimension deviation is symmetrically distributed. The abnormal values of height deviation are concentrated on the side of the larger value, and the median deviation tends to the upper quartile, indicating that the height dimension deviation is right-skewed distribution. Although there are outliers on both sides of the length deviation, the box chart is symmetrical about the median, and the proportion of the outliers in the upper and lower quartiles is consistent; therefore, it can be considered that the length deviation is symmetrical.

4.3. Dimensional Deviation Distribution of Prefabricated Wall Panel Components. Calculating the difference between the measured data and the design value of the prefabricated wall

panel, the sample size is 371, which is divided into 20 groups. The histogram is drawn with the deviation value as the abscissa and the measured frequency as the ordinate, as shown in Figure 10. As seen in Figure 10, the dimensional deviation distribution histogram for the wall panel decreases from the center to the sides, and it can be seen from Figure 10 that the dimensional deviation distribution histogram of the wall panel is a normal distribution.

The height deviation size of prefabricated wall panels is mainly concentrated in the range of $-5-0$ mm, the most frequent deviation is $-3.5-2.8$ mm, the average is -2.39 mm, and the variance is 2.91. The width deviation mainly concentrated in the range of $-3-0.5$ mm, the most frequent deviation was $-1-0.6$ mm, the mean value was -1.13 mm, and the variance was 1.57. The thickness deviation is mainly concentrated in the

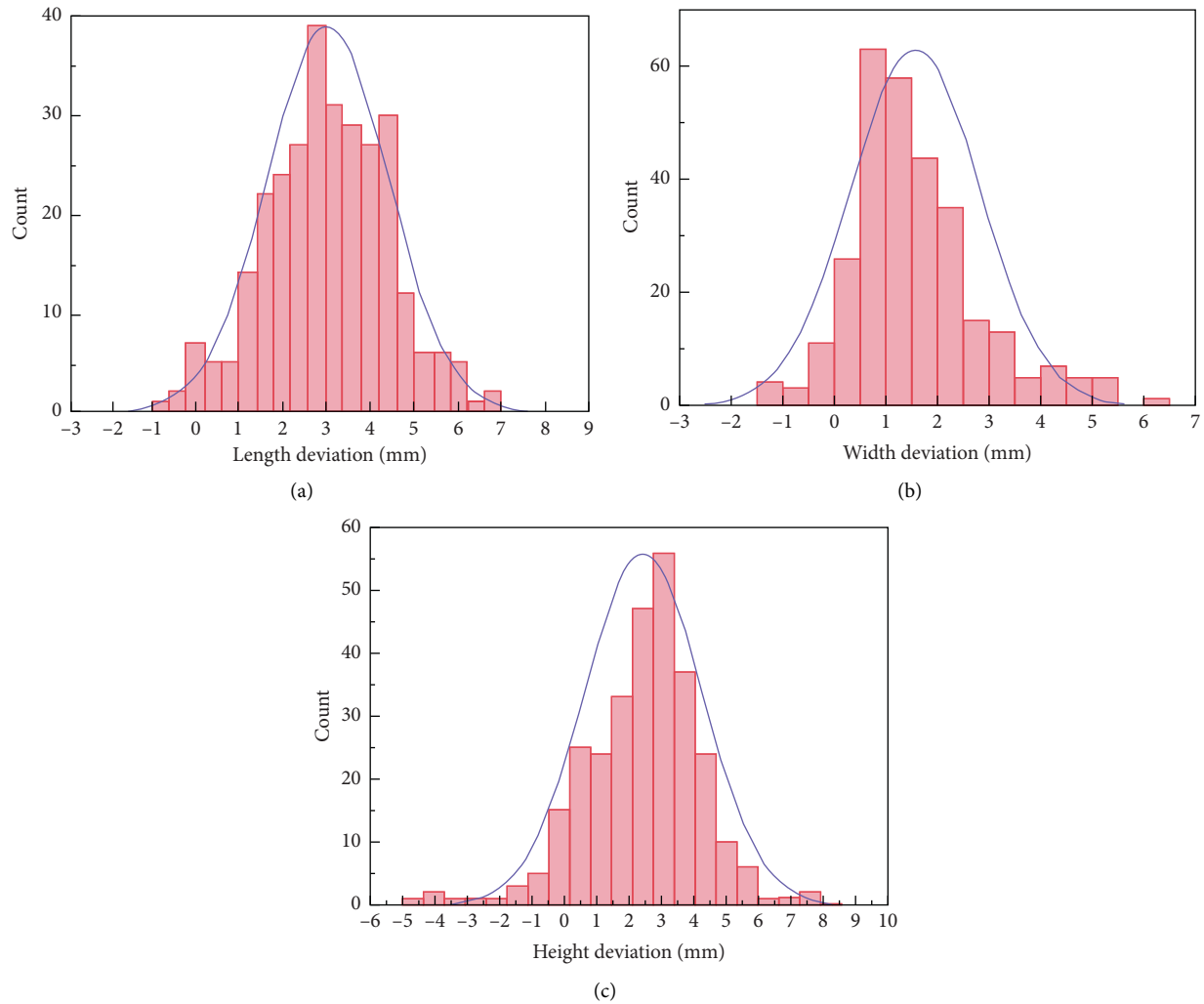


FIGURE 6: Histogram of beam member size deviation distribution. (a) Statistical histogram of beam length deviation. (b) Statistical histogram of beam width deviation. (c) Statistical histogram of beam height deviation.

range of -1.5 – 4 mm, the most frequent deviation is 1.0 – 1.6 mm, the average is 1.81 mm, and the variance is 4.13 .

The maximum, minimum, upper quartile, lower quartile, and median of the size deviations are compared using box plots, as shown in Figure 11. The characteristic values of the three-dimensional dimensional deviation of prefabricated wall panel components are all significantly different. The characteristic values of the wallboard size deviation are significantly different. The height and width deviations are basically negative deviations, and the thickness deviations are basically positive deviations. The abnormal values of height and width deviations are concentrated on the larger value side, and the median deviation tends to the upper quartile indicating that the height dimension deviation and width dimension deviation are right-skewed distribution, and the thickness deviation abnormal values are concentrated on the smaller value side, which means that the thickness deviation distribution is left-skewed distribution.

4.4. Dimensional Deviation Distribution of Prefabricated Composite Slab Components. Because the thickness direction of the composite slab is cast-in-place concrete, the

thickness of the composite slab is not counted. Calculating the difference between the measured data and the design value of the prefabricated composite slab components, the sample size is 401 , which is divided into 20 groups. The histogram is drawn with the deviation value as the abscissa and the measured frequency as the ordinate, as shown in Figure 12. As seen in Figure 12, the dimensional deviation distribution histogram for composite slab decreases from the center to the sides, and it can be seen from Figure 12 that the dimensional deviation distribution histogram of the composite slab is a normal distribution.

The length deviation size of the composite slab is mainly concentrated in the range of -5 – 0 mm, the most frequent deviation is -2.6 – 2.0 mm, the mean value is -2.44 mm, and the variance is 2.21 ; The width deviation mainly concentrated in the range of 0 – 3 mm, the most frequent deviation is -2.5 – 2.0 mm, the mean value is 1.34 mm, and the variance is 1.21 .

The maximum, minimum, upper quartile, lower quartile, and median of the size deviations are compared using box plots, as shown in Figure 13. The dimensional

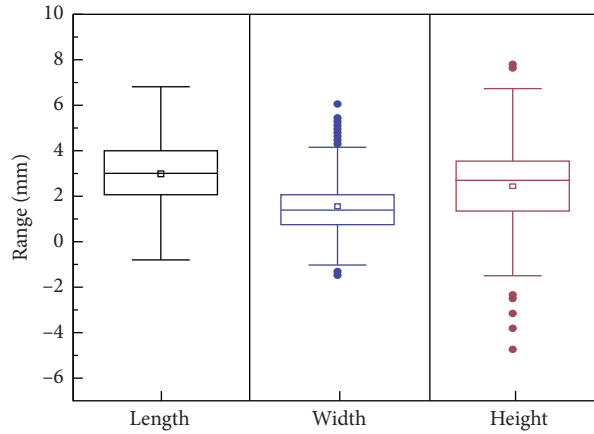


FIGURE 7: Box plot of beam size deviation distribution.

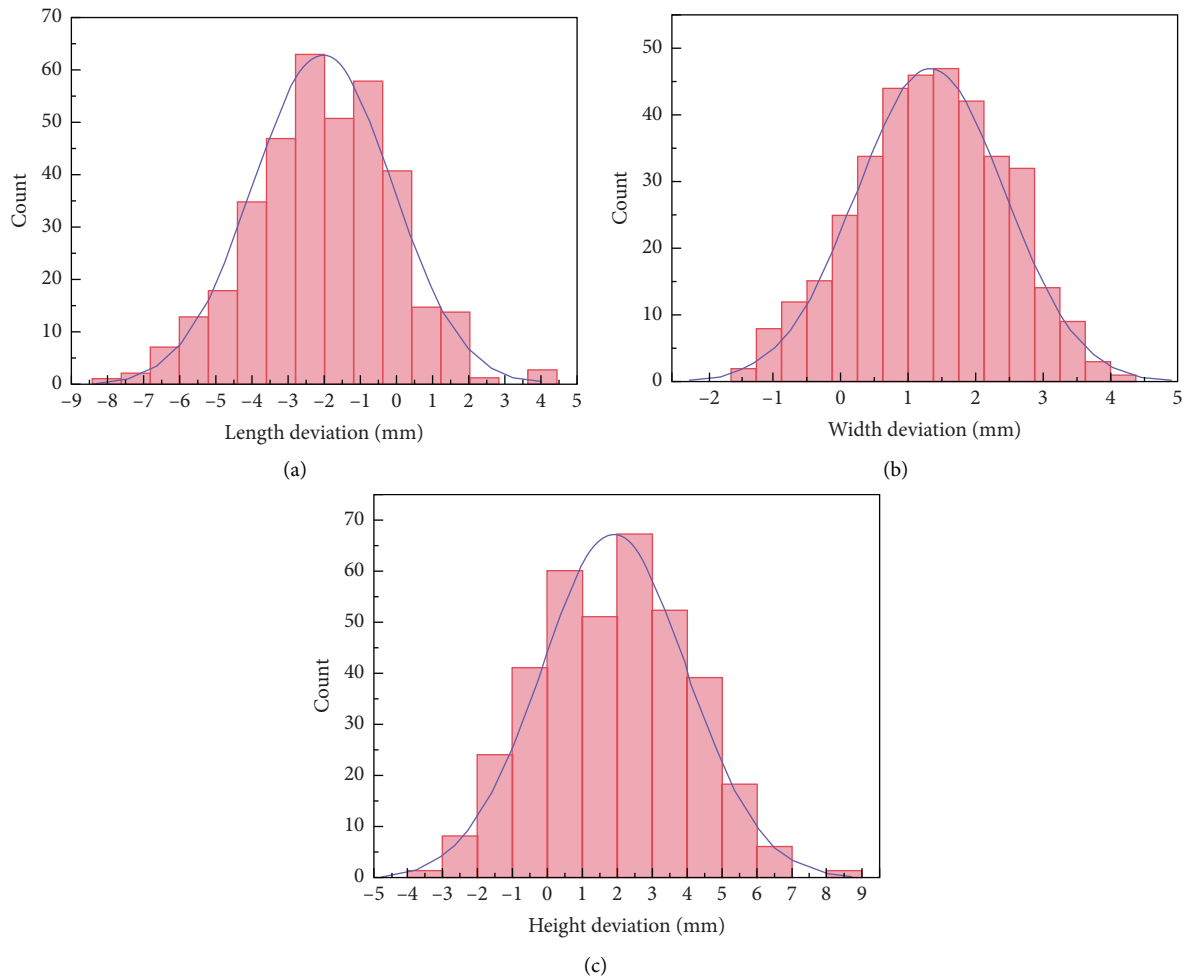


FIGURE 8: Histogram of column member size deviation distribution. (a) Statistical histogram of column length deviation. (b) Statistical histogram of column width deviation. (c) Statistical histogram of column height deviation.

deviation characteristic values of the composite slab slabs are all significantly different, and the dimensional deviations are basically negative deviations, there are abnormal values in the length and height deviation. The length deviation abnormal value is concentrated on the

larger value side, which means that the length deviation distribution is skewed to the left-skewed distribution, and the width deviation abnormal values are concentrated on the side of the smaller value, which means that the width deviation distribution is right-skewed distribution.

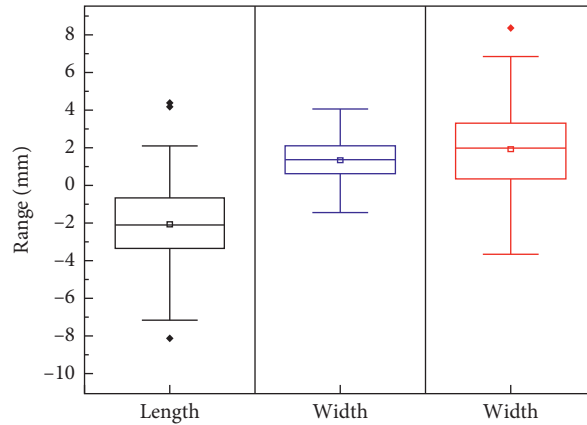


FIGURE 9: Box plot of column size deviation distribution.

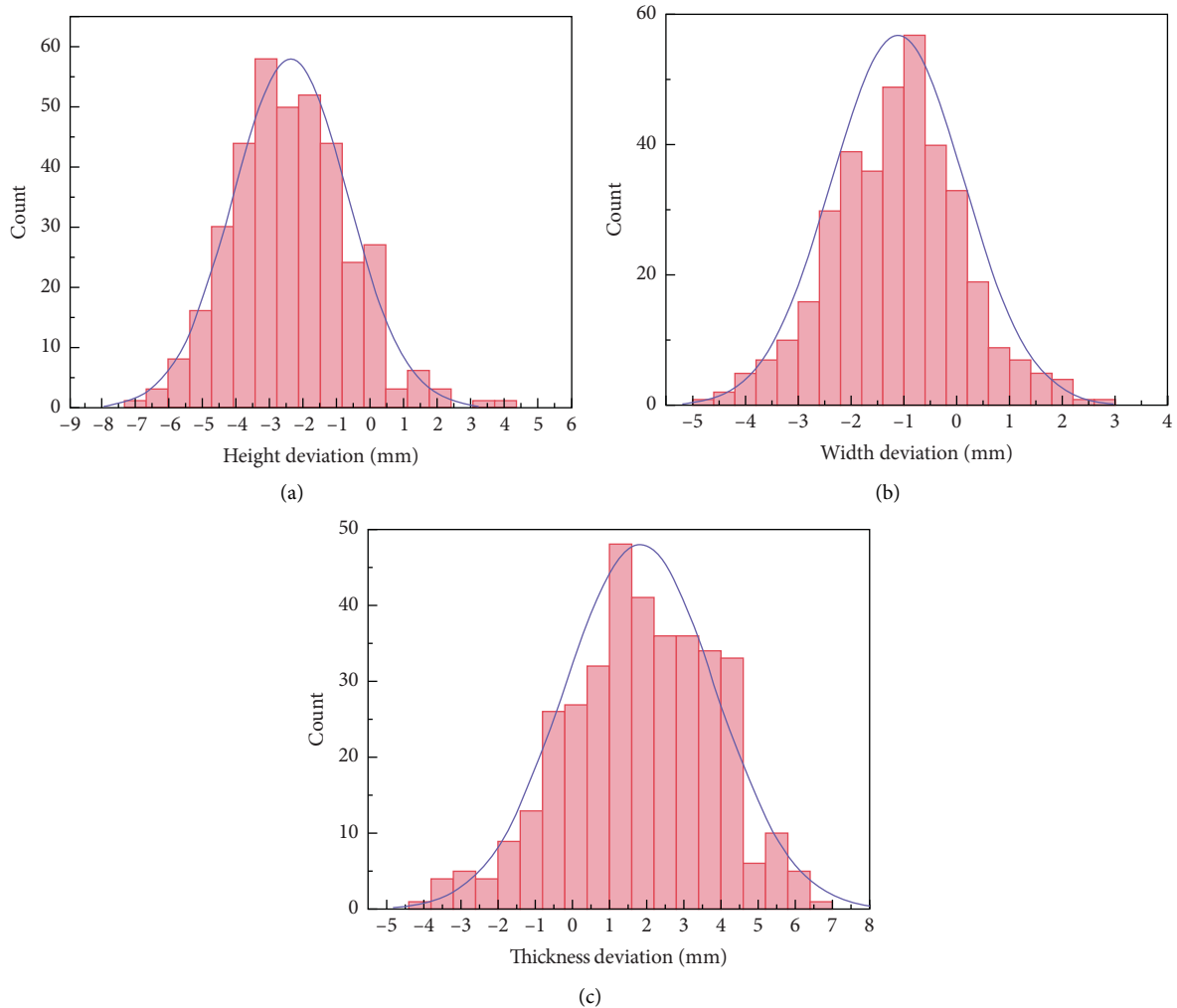


FIGURE 10: Histogram of wall panel size deviation. (a) Statistical histogram of wall panel height deviation. (b) Statistical histogram of wall panel width deviation. (c) Statistical histogram of wall panel thickness deviation.

4.5. *Frequency Distribution Inspection of Dimensional Deviation.* In statistics, the Kolmogorov–Smirnov test (KS test) [18, 19] is used to test whether two empirical

distributions are the same or whether one empirical distribution is different from another ideal distribution. Assuming that X_1, X_2, \dots, X_N and Y_1, Y_2, \dots, Y_N are two

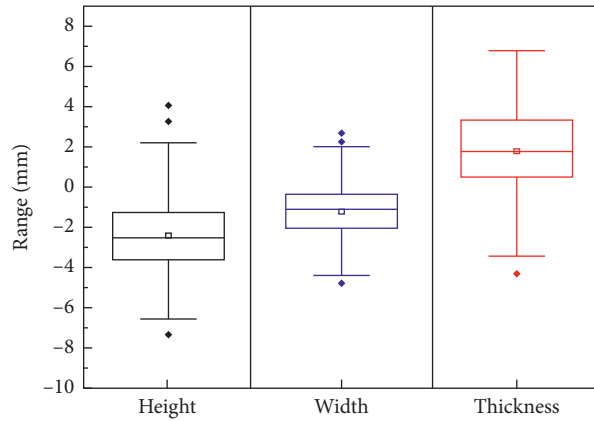


FIGURE 11: Box plot of wall panel size deviation distribution.

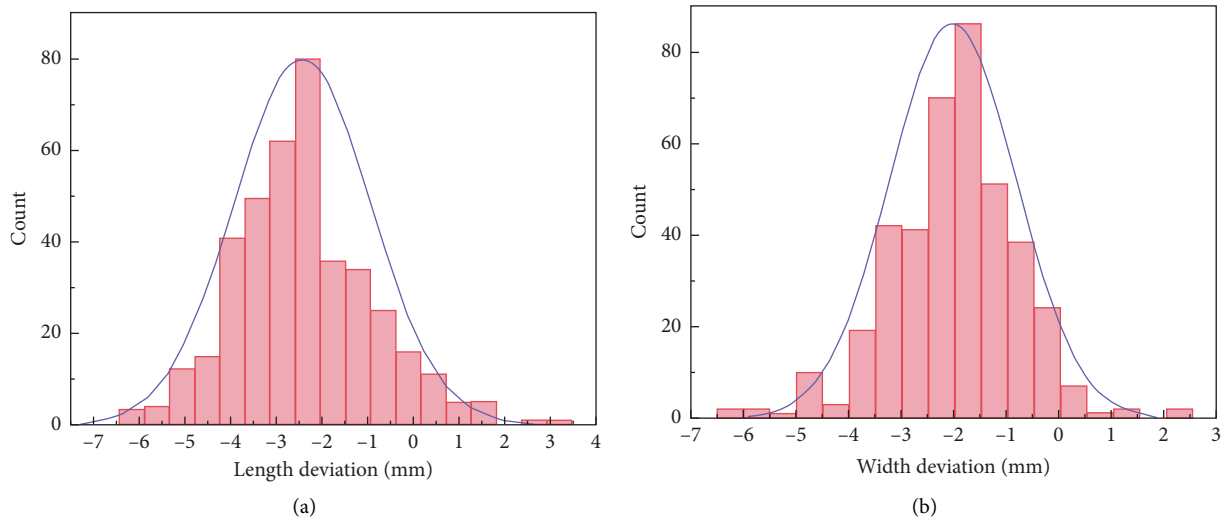


FIGURE 12: Histogram of composite slab size deviation distribution. (a) Statistical histogram of length deviation of composite slab. (b) Statistical histogram of width deviation of the composite slab.

different samples, $K-S$ can be used to test whether the two samples belong to the same population, and it depends on the value of the test statistic D . The D value is defined as

$$D = \max|S_X(x) - S_Y(x)|. \quad (6)$$

Among them, $S_X(x)$ and $S_Y(x)$ are the cumulative distribution functions of the two samples, respectively. If the value of D is small, it can be judged that the two samples are from the same population. Taking into account that the versatility of the normal probability distribution to random variables and the general significance of Weibull's theory in analyzing the size effect of concrete [20], these two probability statistical distributions are compared with the size deviation values of fabricated components. By using Matlab, the normal distribution function $y = \text{normrnd}(\bar{X}, \sigma, m, n)$ is used to generate a sample that conforms to the normal distribution with the average value \bar{X} and the standard deviation σ . The Weibull distribution function $y = \text{weibrnd}(\bar{X}, a, b)$ is used to generate a sample that conforms to the Weibull distribution with \bar{X} as the average, where a is the scale parameter and b is the shape parameter.

Figures 14–17 are the curve fitting diagram of the cumulative distribution function of different component size deviations. The Weibull cumulative distribution function curve and the normal distribution function curve are similar to the component size deviation cumulative distribution function curve. Intuitively, the length deviation cumulative distribution and the width size deviation fit well with the normal distribution, but are significantly different with the Weibull distribution. In the deviation distribution of wallboard height and wallboard thickness, the Weibull cumulative distribution and the normal cumulative distribution function curve are in good agreement. To accurately compare and determine the best distribution, it is necessary to further calculate the maximum absolute deviation among the theoretical distribution function, the empirical distribution function, and the probability P of making the first type of error. The specific results are shown in Table 4. Among them, D is calculated according to equation (6), and the value of P is calculated under the hypothesis of the Kolmogorov distribution of the limit distribution of D .

The KS test results of size deviation frequency, normal distribution, and Weibull distribution are shown in Table 7.

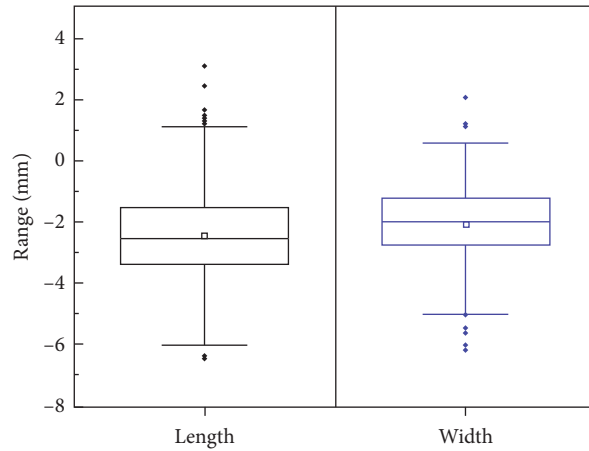


FIGURE 13: Box plot of composite slab size deviation distribution.

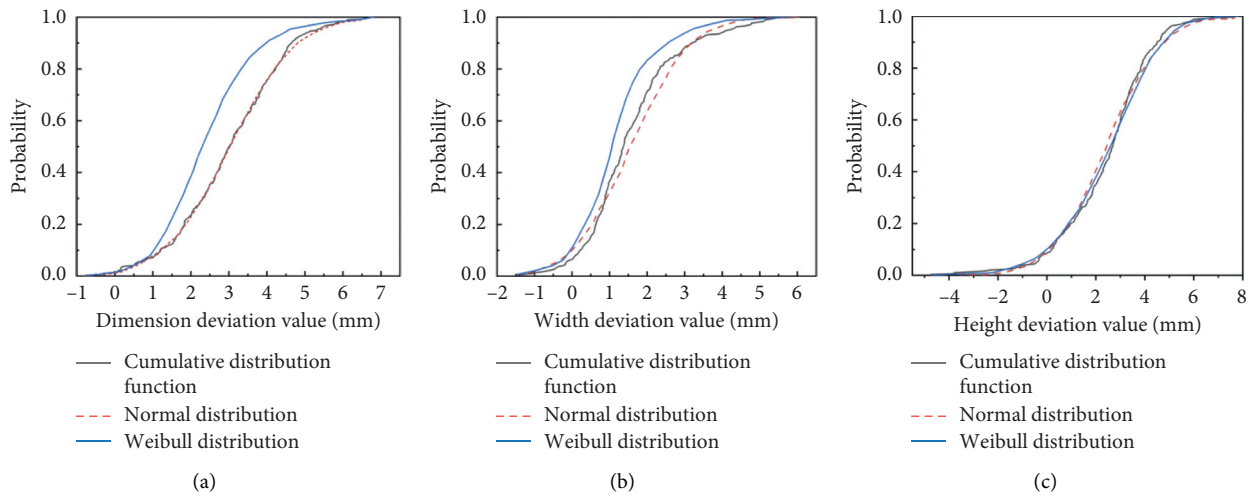


FIGURE 14: The cumulative distribution function of beam size deviation. (a) The cumulative distribution function of beam length deviation. (b) The cumulative distribution function of beam width deviation. (c) The cumulative distribution function of beam height deviation.

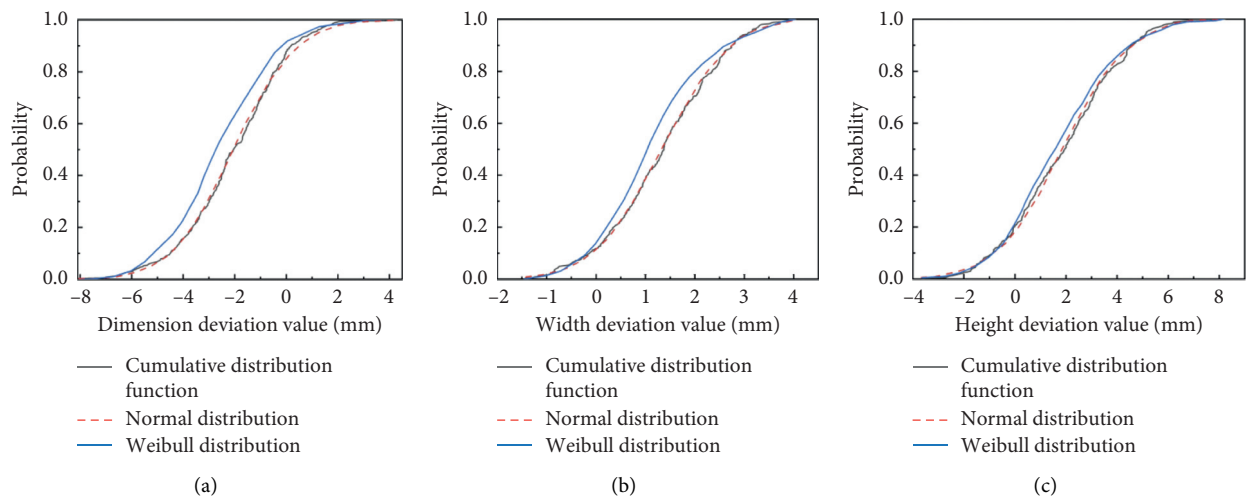


FIGURE 15: The cumulative distribution function of column size deviation. (a) The cumulative distribution function of column length deviation. (b) The cumulative distribution function of column width deviation. (c) The cumulative distribution function of column height deviation.

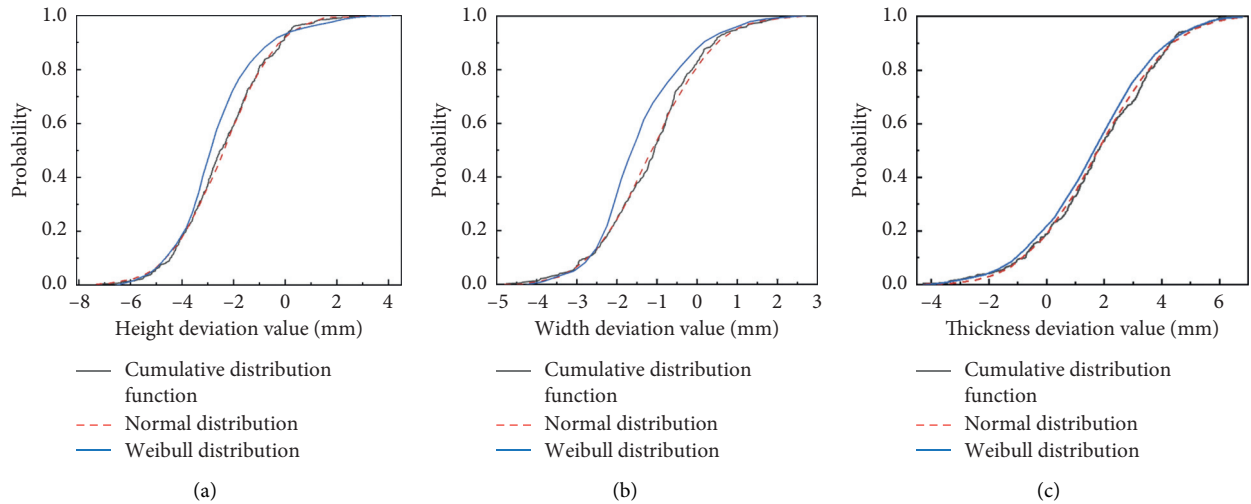


FIGURE 16: The cumulative distribution function of sample value of wall panel size deviation. (a) The cumulative distribution function of wall panel height deviation. (b) The cumulative distribution function of wall panel width deviation. (c) The cumulative distribution function of wall panel thickness deviation.

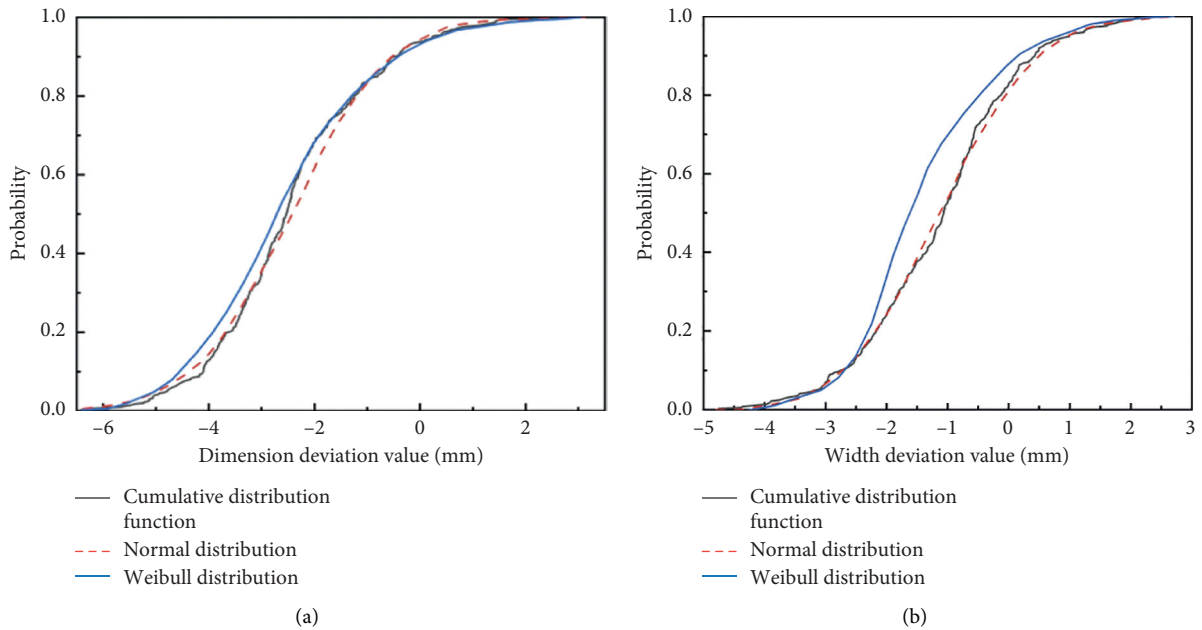


FIGURE 17: The cumulative distribution function of sample value deviation of superimposed slab size. (a) The cumulative distribution function of length deviation of composite. (b) The cumulative distribution function of height deviation of composite slab.

It can be seen from Table 7 that, in the two distributions, the maximum absolute value of the normal distribution D is the smallest, and the P value of the normal distribution is the

smallest. If the significance is taken at the level of 0.05, the fitting effects of normal distribution and Weibull distribution are both acceptable, and it is obvious that the former is

TABLE 7: KS inspection result of size deviation.

Prefabricated components	Project	Normal distribution		Weibull distribution	
		<i>D</i>	<i>P</i>	<i>D</i>	<i>P</i>
Prefabricated beam member	Length	0.038	0.592	0.213	0.039
	Width	0.177	0.187	0.098	0.028
	Height	0.044	0.331	0.052	0.311
Prefabricated column member	Length	0.073	0.429	0.189	0.033
	Width	0.098	0.373	0.167	0.048
	Height	0.031	0.333	0.094	0.309
Prefabricated wall panel components	Height	0.055	0.419	0.193	0.037
	Width	0.061	0.380	0.313	0.018
	Thickness	0.035	0.377	0.051	0.293
Prefabricated composite slab*	Length	0.087	0.199	0.199	0.038
	Width	0.019	0.211	0.227	0.031

*Since the thickness direction of the prefabricated composite slab is cast-in-place concrete, it was not counted.

better. It shows that the dimensional deviation of fabricated components conforms to the null hypothesis and normal distribution.

5. Conclusions

In this paper, three-dimensional surveys of more than 1,400 prefabricated components in China were carried out, and the geometric parameters of the components were statistically analyzed using probability and statistics methods, hence obtaining the following conclusions:

- (1) The range of the geometric parameter uncertainty random variable of the same component size produced by different enterprises shows little difference and the fluctuation is small, ranging from 0.99 to 1.02. The geometric parameter uncertainty coefficient of variation is below 0.0093, and the component size deviation variability is small.
- (2) The significance of the correlation degree of the prefabricated component size deviation is greater than the level of 0.05, and there is no significant difference among the size deviations of similar components produced by different companies.
- (3) The histogram of the frequency distribution of prefabricated component size deviation decreases from the center to both sides, and the fitting curve has only one peak. The geometric uncertainties of the transverse dimensions of the components are all less than 1, which tend to be negative deviations, and the geometric uncertainties of the longitudinal dimensions are all greater than 1, which tend to be positive deviations.
- (4) The size deviation of fabricated components does not refuse to obey the normal distribution and the Weibull distribution, but it is more inclined to obey the normal distribution.

Data Availability

The [xls] data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors gratefully acknowledge the financial support for this research by the applied basic research program of China's 13th Five-Year Plan (2016YFC0701705-1).

References

- [1] T. Song, *Study on quality control of prefabricated concrete structure*, Ph.D. dissertation, Xi'an University of Architecture and Technology, Xian, China, 2017.
- [2] C. T. Milberg and I. D. Tommelein, "Methods for managing tolerance compatibility: Windows in cast-in-place concrete," *Journal of Construction Engineering and Management*, vol. 146, no. 2, Article ID 04019105, 2020.
- [3] S. Talebi, "Improvement of dimensional tolerance management in construction," Ph.D. dissertation, University of Huddersfield, Huddersfield, WY, UK, 2019.
- [4] C. Liu, S. Shirowzhan, S. M. E. Sepasgozar, and A. Kaboli, "Evaluation of classical operators and fuzzy logic algorithms for edge detection of panels at exterior cladding of buildings," *Buildings*, vol. 9, no. 2, p. 40, 2019.
- [5] S. Talebi, L. Koskela, P. Tzortzopoulos, M. Kagioglou, and A. Krulikowski, "Deploying geometric dimensioning and tolerancing in construction," *Buildings*, vol. 10, no. 4, p. 62, 2020.
- [6] DIN18203-1997, *Tolerances for Building Part1: Prefabricated Ordinary, Reinforced and Prestressed Concrete Components*, Deutsches Institut fur Normung E.V. (DIN), Berlin, Germany, 1997.
- [7] PCI MNL-116, *Manual for Quality Control for Plants and Production of Prefabricated and Prestressed Concrete Products*, Prestressed Concrete Institute, Chicago, IL, USA, 2nd edition, 1977.
- [8] ACI 117M-06 and ACI Committee 117, *Specifications for Tolerances for Concrete Construction and Materials and Commentary.the United States*, American Concrete Institute, Farmington Hills, MI, USA, 2006.
- [9] GB50204-2015, *Code for Quality Acceptance of Concrete Structure Construction*, China Architecture & Building Press, Beijing, China, 2015.

- [10] GB/T51231-2016, *Technical Code for Prefabricated*, China Architecture & Building Press, Beijing, China, 2016.
- [11] K. Liu, Y. Z. Wang, and G. Liu, "The statistic and analysis of the dimension deviation of cast-in-situ concrete," *Construction Technology*, vol. 01, no. 1, pp. 98–100, 2008.
- [12] W. C. Zhang, "The Research on the Control Indicators of Construction Erection Variation of Prefabricated Concrete Structure," Ph.D. dissertation, North China University of Science and Technology, Tangshan, China, 2019.
- [13] Research Group on Reliability of Reinforced Concrete Structural Members, "Investigation and statistical analysis of the dimensional tolerances reinforced concrete structural members," *Journal of Building Structures*, vol. 6, no. 4, pp. 2–9, 1985.
- [14] H. X. Mao and W. Z. Shi, "New methodology of representing the positional error of non-point features in GIS," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 37, pp. 1503–1508, 2008.
- [15] M. X. Zhang, "Investigation and study on constructional quality control of concrete structure in Situ," Ph.D. dissertation, Zhejiang University, Hangzhou, Zhejiang, China, 2004.
- [16] X. P. Hu, "Investigation and statistical analysis on resistance of building structural members and reliability analysis on residential buildings," Ph.D. dissertation, Xi'an University of Architecture and Technology, Xian, China, 2005.
- [17] X. C. Li, "Investigation and statistical analysis on resistance of reinforced concrete structure in Xianyang region," Ph.D. Dissertation, Xi'an University of Architecture and Technology, Xian, China, 2015.
- [18] E. F. Acar and L. Sun, "A generalized kruskal-wallis test incorporating group uncertainty with application to genetic association studies," *Biometrics*, vol. 69, no. 2, pp. 427–435, 2013.
- [19] L. D. Ong and P. C. Leclare, "The Kolmogorov–Smirnov test for the log-normality of sample cumulative frequency distributions," *Health Physics*, vol. 14, no. 4, p. 376, 1968.
- [20] A. Ghasemi and S. Zahediasl, "Normality tests for statistical analysis: a guide for non-statisticians," *International Journal of Endocrinology and Metabolism*, vol. 10, no. 2, pp. 486–489, 2012.