

Journal of Healthcare Engineering

# Computation Methods for Biomedical Information Analysis

Lead Guest Editor: Yong Xia

Guest Editors: Weidong Cai, Xiaofeng Yang, and Shanshan Wang





---

# **Computation Methods for Biomedical Information Analysis**

Journal of Healthcare Engineering

---

## **Computation Methods for Biomedical Information Analysis**

Lead Guest Editor: Yong Xia

Guest Editors: Weidong Cai, Xiaofeng Yang, and Shanshan Wang



---

Copyright © 2018 Hindawi. All rights reserved.

This is a special issue published in “Journal of Healthcare 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.

## Editorial Board

Saverio Affatato, Italy  
Francesca Apollonio, Italy  
Hasan Ayaz, USA  
William Bertucci, France  
Patrick Boissy, Canada  
Hongwei Chen, USA  
Daniel H.K. Chow, Hong Kong  
Gianluca Ciardelli, Italy  
Elena De Momi, Italy  
Costantino Del Gaudio, Italy  
Tiago H. Falk, Canada  
Mostafa Fatemi, USA  
Jesus Favela, Mexico  
David Dagan Feng, Australia  
Joseph Finkelstein, USA  
Jesus Fontecha, Spain  
Jean-Luc Gennisson, France  
Shahram M. Ghanaati, Germany  
Luca Giancardo, USA  
Antonio Gloria, Italy  
Kheng-Lim Goh, Singapore  
Pedro Gomes, Portugal  
Carlos Gómez, Spain  
Philippe Gorce, France  
Sophia Z. Gu, Australia  
Vincenzo Guarino, Italy  
Valentina Hartwig, Italy

David Hewson, UK  
Andreas H. Hielscher, USA  
Ernesto Iadanza, Italy  
Norio Iriguchi, Japan  
Jingfeng Jiang, USA  
Zhongwei Jiang, Japan  
Rashed Karim, UK  
Pasi A. Karjalainen, Finland  
Chandan Karmakar, Australia  
John S. Katsanis, Greece  
Terry K.K. Koo, USA  
Panagiotis Kosmas, UK  
Michel Labrosse, Canada  
Jui-Yang Lai, Taiwan  
Xiang Li, USA  
Feng-Huei Lin, Taiwan  
Yuan-Pin Lin, Taiwan  
Maria Lindén, Sweden  
Dongfei Liu, Finland  
Francisco Lopez-Valdes, Spain  
Zufu Lu, Australia  
Ilias Maglogiannis, Greece  
Andreas Maier, Germany  
Mehran Moazen, UK  
Rafael Morales, Spain  
David Moratal, Spain  
Angkoon Phinyomark, Canada

Vincenzo Positano, Italy  
Alessandro Ramalli, Italy  
Alessandro Reali, Italy  
Lei Ren, UK  
Jose Joaquin Rieta, Spain  
Emanuele Rizzuto, Italy  
Sébastien Roth, France  
Simo Saarakkala, Finland  
Hélder A. Santos, Finland  
Emiliano Schena, Italy  
Maurizio Schmid, Italy  
Jiann-Shing Shieh, Taiwan  
Tiago H. Silva, Portugal  
Redha Taiar, France  
Jinshan Tang, USA  
Vinoy Thomas, USA  
Ioannis G. Tollis, Greece  
Kazunori Uemura, Japan  
Uche Wejinya, USA  
Damon W.K. Wong, Singapore  
Ying Yang, UK  
Haihong Zhang, Singapore  
Hongbo Zhang, Finland  
Ping Zhou, USA  
Loredana Zollo, Italy

# Contents

## **Computation Methods for Biomedical Information Analysis**

Yong Xia , Weidong Cai, Xiaofeng Yang, and Shanshan Wang  
Editorial (2 pages), Article ID 8683601, Volume 2018 (2018)

## **The Wall Apposition Evaluation for a Mechanical Embolus Retrieval Device**

Xuelian Gu , Yongxiang Qi, and Arthur G. Erdman   
Research Article (8 pages), Article ID 9592513, Volume 2018 (2018)

## **A Filtering Method for Identification of Significant Target mRNAs of Coexpressed and Differentially Expressed MicroRNA Clusters**

Su Yeon Lee, Soo-Yong Shin, Young Jo Yoon, and Yu Rang Park   
Research Article (6 pages), Article ID 4932904, Volume 2018 (2018)

## **Comparative Study on Automated Cell Nuclei Segmentation Methods for Cytology Pleural Effusion Images**

Khin Yadanar Win , Somsak Choomchuay, Kazuhiko Hamamoto, and Manasanan Raveesunthornkiat   
Research Article (14 pages), Article ID 9240389, Volume 2018 (2018)

## **The Cell Research Trends of Asthma: A Stem Frequency Analysis of the Literature**

Wenchao Tang , Yi Shang , Bin Xiao , Peitong Wen , Ruoyun Lyu , and Ke Ning   
Research Article (10 pages), Article ID 9363820, Volume 2018 (2018)

## **Link Prediction Investigation of Dynamic Information Flow in Epilepsy**

Yan He , Fan Yang, Yunli Yu, and Celso Grebogi  
Research Article (13 pages), Article ID 8102597, Volume 2018 (2018)

## **A Method for Tooth Model Reconstruction Based on Integration of Multimodal Images**

Xinwen Zhou, Yangzhou Gan, Jing Xiong, Dongxia Zhang, Qunfei Zhao, and Zeyang Xia   
Research Article (8 pages), Article ID 4950131, Volume 2018 (2018)

## **Bayesian Classification Models for Premature Ventricular Contraction Detection on ECG Traces**

Manuel M. Casas, Roberto L. Avitia , Felix F. Gonzalez-Navarro, Jose A. Cardenas-Haro, and Marco A. Reyna  
Research Article (7 pages), Article ID 2694768, Volume 2018 (2018)

## **Predicting Interactions between Virus and Host Proteins Using Repeat Patterns and Composition of Amino Acids**

Saud Alguwaizani, Byungkyu Park, Xiang Zhou, De-Shuang Huang , and Kyungsook Han   
Research Article (9 pages), Article ID 1391265, Volume 2018 (2018)

## **Breast Mass Detection in Digital Mammogram Based on Gestalt Psychology**

Hongyu Wang , Jun Feng , Qirong Bu, Feihong Liu, Min Zhang, Yu Ren, and Yi Lv  
Research Article (13 pages), Article ID 4015613, Volume 2018 (2018)

## **Organic Boundary Location Based on Color-Texture of Visual Perception in Wireless Capsule Endoscopy Video**

Chengliang Wang , Zhuo Luo , Xiaoqi Liu, Jianying Bai, and Guobin Liao  
Research Article (11 pages), Article ID 3090341, Volume 2018 (2018)

## Editorial

# Computation Methods for Biomedical Information Analysis

**Yong Xia** <sup>1</sup>, **Weidong Cai**<sup>2</sup>, **Xiaofeng Yang**<sup>3</sup> and **Shanshan Wang**<sup>4</sup>

<sup>1</sup>*School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an, China*

<sup>2</sup>*School of Information Technologies, University of Sydney, Sydney, NSW 2006, Australia*

<sup>3</sup>*Department of Radiation Oncology and Winship Cancer Institute, Emory University, Atlanta, GA 30322, USA*

<sup>4</sup>*Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China*

Correspondence should be addressed to Yong Xia; [yxia@nwpu.edu.cn](mailto:yxia@nwpu.edu.cn)

Received 30 September 2018; Accepted 30 September 2018; Published 1 November 2018

Copyright © 2018 Yong Xia 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.

With the advances in computer technology, biomedical data have been growing at an explosive speed, especially the massive accumulation of clinical medical data. It highly raises the requirement of data processing and analysis. Meanwhile, the computation methods can not only help medical professionals and researchers to get rid of the limitations of manual information processing in traditional methods but also provide them with more effective methods to use data. With the application of computation methods to biomedical information analysis, effective ways have been proposed to solve some problems that cannot be solved before.

This special issue provides a platform for medical professionals and researchers to present their ideas about the latest issues and developments in biomedical analysis fields. The papers in this special issue address the development and application of medical image segmentation, classification, detection, modeling, predicting, and literature mining.

The major part of this special issue is a collection of papers on image processing and analysis in various application areas.

The paper “Predicting Interactions between Virus and Host Proteins Using Repeat Patterns and Composition of Amino Acids” uses the repeat patterns and composition of amino acids to predict interactions between virus and host proteins and find new virus-host protein-protein interactions (PPIs) with little information.

The paper “Bayesian Classification Models for Premature Ventricular Contraction Detection on ECG Traces” compares three well-known Bayesian classification algorithms, including the naive Bayes, linear discriminant analysis, and

quadratic discriminant analysis algorithms, in distinguishing the normal heartbeats, premature ventricular contraction beats, and others.

The paper “Organic Boundary Location Based on Color-Texture of Visual Perception in Wireless Capsule Endoscopy Video” combines the color-saliency region detection method and support vector machine (SVM) classifier to promote the efficiency and accuracy of locating the pylorus in wireless capsule endoscopy videos.

The paper “Breast Mass Detection in Digital Mammogram Based on Gestalt Psychology” incorporates visual perception properties described by the Gestalt psychology framework into breast mass detection in a digital mammogram to achieve good performance on both the Mammographic Image Analysis Society (MIAS) data set and Digital Database for Screening Mammography (DDSM) data set.

The paper “A Filtering Method for Identification of Significant Target mRNAs of Coexpressed and Differentially Expressed MicroRNA Clusters” devises to use hypergeometric distributions to identify significant miRNA target genes from an extensive list of predicted miRNA target gene relationships.

The paper “A Method for Tooth Model Reconstruction Based on Integration of Multimodal Images” develops a tooth model reconstruction method based on integration of computed tomography (CT) images and laser scan images, which can generate a tooth model with a more accurate crown and can obtain a complete tooth model at any stage of orthodontic treatment by using one CT scan at the pretreatment stage and one laser scan at that stage to avoid multiple CT scans.

The paper “Link Prediction Investigation of Dynamic Information Flow in Epilepsy” adapts link prediction for proposing a novel workflow, which detects the seizure occurrence and monitors the total seizure course for quantifying seizure dynamics and uncovers pathological mechanisms of epilepsy.

The paper “Comparative Study on Automated Cell Nuclei Segmentation Methods for Cytology Pleural Effusion Images” presents a comparative study on 12 automated cell nuclei segmentation methods, finding that the segmentation performances of the Otsu, k-means, mean shift, Chan–Vese, and graph cut methods outperform the others, which will be useful for current and potential future studies on cytology images of pleural effusion.

This special issue also has two papers, which describe medical data retrieval systems.

The paper “The Cell Research Trends of Asthma: A Stem Frequency Analysis of the Literature” summarizes the asthma literature indexed in the medical literature analysis and retrieval system online (MEDLINE) and explores the history to present trends of asthma cell research by stem frequency ranking to forecast the prospect of the future work.

The paper “The Wall Apposition Evaluation for a Mechanical Embolus Retrieval Device” presents a computational evaluation approach to the wall apposition of a cerebral mechanical emboli retrieval device (MERD) and provides references and suggestions for further research and work.

Although it is impossible to comprehensively cover the growing field of computation methods for biomedical information analysis in a special issue like this, we believe that this special issue presents a set of state-of-the-art computational methods and their potentials in the biomedical information analysis and related domains. We hope this issue would inspire further research on the aforementioned topics.

## **Conflicts of Interest**

The editors declare that they have no conflicts of interest.

## **Acknowledgments**

As guest editors of this special issue, we would like to thank all the authors who have submitted their contributions to this special issue; without their interests and collaborations, this issue would not have been possible. We would also like to thank all referees for their rigorous reviews with insightful and constructive comments that have definitely improved the quality of this issue.

*Yong Xia  
Weidong Cai  
Xiaofeng Yang  
Shanshan Wang*

## Research Article

# The Wall Apposition Evaluation for a Mechanical Embolus Retrieval Device

Xuelian Gu <sup>1</sup>, Yongxiang Qi,<sup>2</sup> and Arthur G. Erdman <sup>3</sup>

<sup>1</sup>Shanghai Institute for Minimally Invasive Therapy, University of Shanghai for Science and Technology, 516 Jungong Road, Shanghai 200093, China

<sup>2</sup>MicroPort Endovascular (Shanghai) Co., Ltd., 3399 Kangxin Rd., Shanghai 201318, China

<sup>3</sup>Department of Mechanical Engineering, University of Minnesota, 111 Church St. SE, Minneapolis, MN 55455, USA

Correspondence should be addressed to Xuelian Gu; [guxuelian@usst.edu.cn](mailto:guxuelian@usst.edu.cn) and Arthur G. Erdman; [agerdman@umn.edu](mailto:agerdman@umn.edu)

Received 4 November 2017; Revised 2 April 2018; Accepted 3 May 2018; Published 25 September 2018

Academic Editor: Shanshan Wang

Copyright © 2018 Xuelian Gu 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.

A computational evaluation approach to the wall apposition of a cerebral mechanical emboli retrieval device (MERD) is presented. The typical enclosed multilattice structure, manufactured from the thin-walled Nitinol tube, consists of repeated “V”-shaped unit cells. During interventional thrombectomy, the MERD system is delivered inside an artery stenosis segment to capture emboli and restore cerebral blood flow. The wall apposition, which deteriorates during embolus capture, occurs during system migration along the tortuous intracranial vessel. The commercial finite element analysis (FEA) solver ABAQUS 6.10 Standard and user subroutine (UMAT/Nitinol) are utilized to study the ability to remain in close contact with the curved vessel wall during migration. In this numerical analysis, the influence of the contacting interference loadings on structure deformation and strain field distribution is obtained and analyzed. The results indicate that the middle segment of the MERD seriously contracts or collapses inside the curved vessel. In addition, the peak strain is in the apex flow-prone region and maintains at the safe range.

## 1. Introduction

Stroke is a type of brain function disease which is due to the disturbance of blood supply to the brain tissue. Stroke has already become the second leading cause of death in the world, and approximately 88% is acute ischemic stroke (AIS) [1]. It is reported that the number of deep venous thrombosis cases has increased to 250,000 per year in the United States, and the ratio of morbidity or mortality is also fairly high [2]. During the past decade, mechanical embolus retrieval devices (MERDs) have been extensively used as an interventional neurovascular device to treat AIS [3–6]. The MERD is designed to capture the embolus or thrombus, scaffold the afflicted artery, and restore blood flow. This paper reports the analysis of the MERD which is fabricated from the nickel-titanium alloy thin-walled tube by laser cutting. After electropolishing and heat treatment for the shape setting, the embolus capture device is shaped like an enclosed tubular multilattice structural “funnel.” MERD structures vary in complexity from a 3D helical coil to a laser-cut nickel-titanium (Nitinol) tube. So far, several embolus retrieval devices have been used in clinics, such as the Merci

Retrieval system (Concentric Medical, Mountain View, California, USA), the Penumbra Aspiration system and the Stent Retrievers, and others [5, 6]. The latest generation of mechanical thrombectomy devices is the ‘stent retriever’ family: Solitaire (ev3 Endovascular, Plymouth, Minnesota, USA), Trevo Pro (Stryker Neurovascular, Kalamazoo Michigan, USA), and arguably the Penumbra 3D separator (Penumbra, Alameda, California, USA) [7–9].

In clinical application, the MERD works according to the following steps:

- (1) Crimp the stent into a catheter and delivery to the stenosed (blocked) vessel, as shown in Figure 1(a)
- (2) Deploy to capture the embolus while the outer catheter is removed, as shown in Figures 1(b) and 1(c)
- (3) Migrate the embolus removal device and extract tissue from the patient’s body, as shown in Figure 1(d).

The cerebral blood vessels have a relatively small profile and often small radius of curvature in the anatomy. After the MERD has been inserted and has migrated inside the

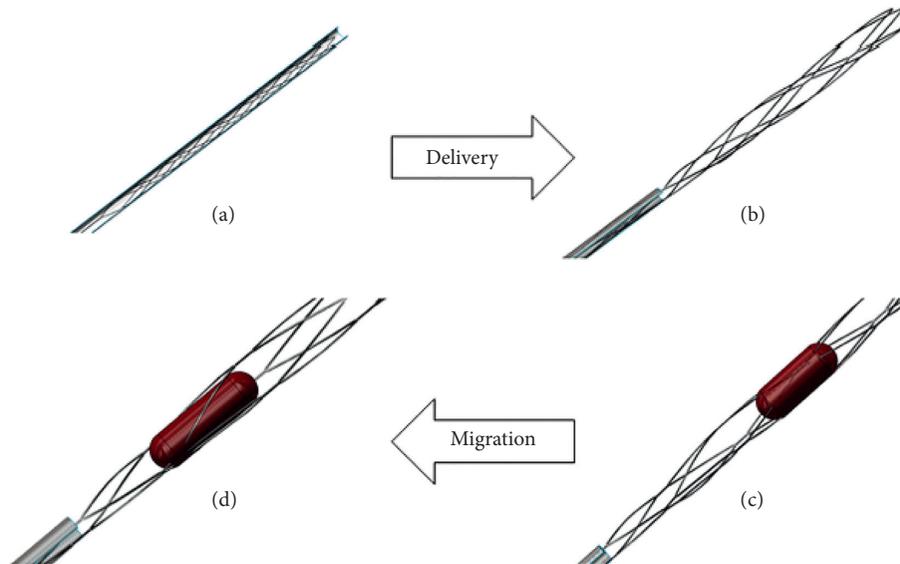


FIGURE 1: The numerical workflow of the MERD. (a) Crimp. (b) Deploy. (c) Extract. (d) Capture.

tortuous cerebral vessel, the embolus retrieval performance is of great importance [3, 5]. The wall apposition, which describes the stent or MEDR's ability to remain in close contact with the adjacent vessel wall [3], is a significant mechanical requirement of a MERD system while it is deployed in a curved vessel and will affect the effectiveness and accuracy of embolus retrieval.

So far, few researches have been published to study the mechanical behavior of a thrombectomy device with a numerical modeling method. However, the mechanical performance of multilattice tubular devices, such as vascular stents, has been studied already [3, 10–15]. Krischek et al. proposed the mechanical features for a MERD, including the radial force, wall apposition, conformability, and Gator backing (Gator backing describes a stent's tendency to flair its struts outward, forming protrusions, when the stent is placed around a bend) and carried out experimental tests [3]. Kleinstreuer et al. simulated crimping, deployment, and cyclic-loading procedures. A high-cycle fatigue prediction method for mean strain/alternating strain in Nitinol material was established in [10]. Kate et al. compared nonstent retriever and stent retriever mechanical thrombectomy devices for the endovascular treatment of acute ischemic stroke. Stent retriever mechanical thrombectomy devices achieve higher recanalization rates than nonstent retriever devices in acute ischemic stroke with improved clinical and radiographic outcomes and safety [11]. Azaouzi et al. analyzed the deployment of a self-expanding stent inside an artery by FEA, and the results could be used to assess the impact of the stent on the artery and the influence of the artery on the deformation field within the stent [12]. Gu et al. presented a numerical analysis of a semienclosed tubular mechanical embolus retrieval device (MERD) for the treatment of AIS, and the FEA methodology is used to evaluate mechanical performance and provide suggestions for optimizing the geometric design [13]. Wu et al. generated a finite element (FE) model of a vascular stent with tetra-elements.

The axial flexibility of the stent is studied by applying moment in [14]. Grogan et al. designed CoCr stents and compared the mechanical properties of the deployment, radial force, longitudinal resistance, and flexibility. The difference between a rigid expanding tool and a polymer balloon is demonstrated in [15].

After being delivered to the targeted artery wall, Nitinol stents expand and contact with the artery wall. The tubular stent is designed to be embedded in the stenosed blood vessel segment to avoid movement. The multilattice MERD, however, is manipulated so that it migrates inside the blood lumen to capture and retrieve the plaque blockage, as shown in Figures 1(c) and 1(d). Poor wall apposition inside tortuous vessels may result in undesired clinical accidents and even failure to retrieve plaque. So far, the numerical simulation process of embolus retrieval in tortuous vessel walls has not been reported. In interventional thrombectomy, the Nitinol MERD functions as a disposable surgical instrument rather than a permanent implant. Prediction of structural failure depends on the peak strain value. Finally, the unique design of the distal and proximal structure constructed is demonstrated. The paper develops a scientific numerical analysis workflow of the MERD for the expansion capability and wall apposition. The analysis method provides a structural optimization scheme for MERD design. In clinical use, an operational guideline for embolus retrieval and migration of the MERD is demonstrated, which will help optimize a specific procedure. In conclusion, the study of the MERD for the wall apposition inside a tortuous blood vessel is provided to improve the device's mechanical capability as well as to aid in surgical training.

In this paper, a typical tubular enclosed mesh-like MERD model is built to study the influence of shape setting and migration. ABAQUS 6.10/Standard (DS SIMULIA, RI, USA) commercial FE code and its user material subroutine (UMAT/Nitinol) are employed to simulate the procedure and the MERD/artery contact interaction mechanism.

Our study presents the following critical measurement of performance: (1) large strain distribution and the highest peak value of maximum principal strain (MPS) in the microstructure and (2) wall apposition performance along the tortuous artery in the macrostructure.

The analysis results are used to assess the safety and efficiency of the MERD. The wall apposition performance evaluation methodology is a scientific numerical method to analyze preexisting structural flaws and offer approaches for design optimization.

**1.1. Methods.** A 3D finite element model has been generated to study the effects of MERD expansion and migration. The numerical simulation processes are achieved with the contact interactions of a MERD/shape-setting cylinder, MERD/expanding cylinder, and MERD/artery at a body temperature of 37°C [10]. The geometry and mesh model of the embolus removing device are shown in Figure 2.

In the numerical analysis, the MERD is assumed to be a homogeneous isotropic incompressible deformed body in the absence of residual stress. The shape-setting cylinder and expanding cylinder are designed to be a semirigid movable cylinder shell. The tortuous cerebral vessel is modeled as a “C”-shaped discrete rigid shell. The Nitinol material constitutive model is characterized by the ABAQUS 6.10 software (UMAT/Nitinol) user subroutine. Figure 3 provides the stress-strain and stress-temperature curves for the template of Nitinol alloy based on ABAQUS software, while Table 1 lists the specified parameters of Nitinol material from the previous research [10]. These parameters demonstrate the distinct mechanical behavior of Nitinol during loading and unloading conditions with specific temperature.

This mesh-like tubular MERD comprises five “column” cells in the axial direction. Each “column” cell consists of four “V”-like wave rings, and each “column” cell is connected with connected bridges circumferentially. The strut thickness, width, original outer diameter, nominal final outer diameter, artery centerline radius, and artery cross-profile diameter are 0.07, 0.07, 2.00, 4.00, 5.00, and 3.00 mm, respectively. The element types C3D8I (three-dimensional eight-node Stress Hex incompatible element), SFM3D4 (three-dimensional four-node quadrilateral surface element), and R3D4 (three-dimensional linear four-node bilinear rigid quadrilateral element) are utilized for the MERD, tortuous vessel, and expanding/shape-setting cylinder, separately. To minimize the influences of mesh density, two layer elements for the wall thickness [16] and 16 elements along the fillet edge are used. The element and node numbers of the MERD FE model are 40,000 and 20,000, respectively. In addition, to form an enclosed thrombus-capturing “cage,” “funnel”-shaped tails are meshed and connected to the MERD’s FE principal body after expansion and annealing. At the end of shape setting, a Nitinol tortuous guiding wire is tied to the “funnel”-shaped head for the pulling/pushing simulation. A flare-shaped tube is connected to the “C”-shaped cerebral vessel. The novel structural design offers smooth and stable surface interaction to avoid computational contact divergence. The boundary conditions for shape setting and migration are described in Figure 4.

In the shape-setting step, a radial outward displacement is imposed on the expanding cylinder to achieve the expanded shape size. Besides, it is also constrained axially and circumferentially, that is, preventing the transitional and rotational movements. Meanwhile, the shape-setting cylinder is restrained in all DOF to avoid offset. A single node of MERD instance is fixed axially to restrict movement. In the migration step, the reference point of a rigid artery is restrained in all DOFs under a global rectangular coordinate system. Negative axial displacement, imposed on the guiding-wire side, is used to pull the MERD along the desired tortuous artery.

The master/slave contact pair algorithm is utilized to build the contact interface between the expanding/shape-setting cylinder and the MERD. Meanwhile, the arterial inner surface is taken as the master surface, and the MERD outer surface is set as the slave surface. The penalty contact method is employed. It is notable that self-contact interaction should also be considered to avoid struts overlapping during the migration step. In addition, a damping factor could be employed discreetly to stabilize the contact-induced vibration behavior, improve convergence, and reduce computational expense in contacting interaction domains.

## 2. Results

**2.1. Validation.** To validate the finite element model and computational algorithm, a diamond-shaped pattern is generated for the numerical simulation and experimental test. The comparison between test and FEA radial forces is used to predict the accuracy and reliability. In the experimental test, the measurement system applies displacement on several metal slices to compress the multilattice pattern. The electrical force sensor (RX500, Machine Solutions Inc.), shown in Figure 5, obtains and reflects the real-time support force value. For the radial force (RF) curve plots, the outer diameter of the MERD is designed as the horizontal  $x$ -axis, while the vertical axis prescribes the balanced force value. In the progress of numerical simulation, the crimping slice is replaced by a rigid removable cylinder to compress the original pattern. The curves for the test and numerical process are shown in Figure 6.

The result indicates that the test and simulation curve match well; that is, all of the numerical outcomes are relatively accurate and could therefore provide a reliable performance analysis and structure optimization report.

**2.2. Shape Setting.** The original laser-cut tubular multilattice structural MERD is expanded by imposing an outward radial displacement on a cylinder surface. Once the final shape is accomplished, the mechanical behavior of cross-profile expansion and axial length shrinkage is presented in Figure 7.

During shape setting, the obtained radial outer diameter ranges from 2.00 to 4.00 mm, and the axial length ranges from 42.9 to 40.0 mm. In contrast to traditional metallic material, the Nitinol material fracture and fatigue failure are strain induced [10]. Structural damage can occur from outright fracture during the expansion step. Figure 7 presents the strain field contour plots of the expanded MERD. It indicates

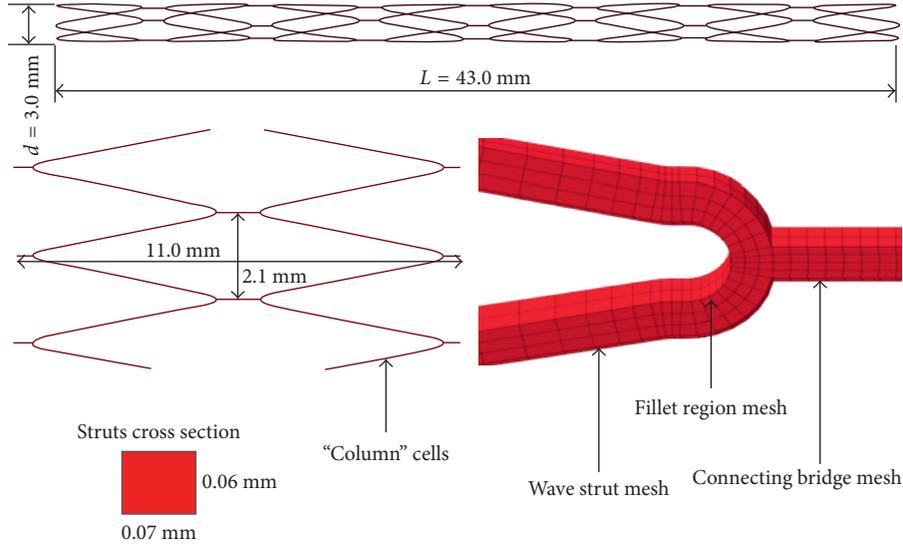


FIGURE 2: Geometry parameters and mesh model of the MERD.

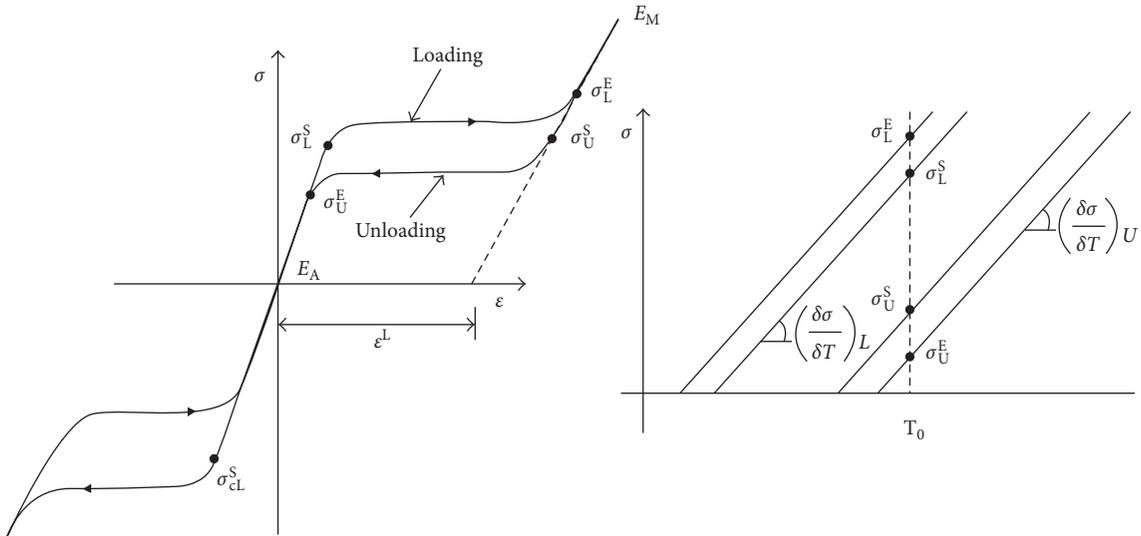


FIGURE 3: Nitinol material properties (from ABAQUS Nitinol UMAT) [10].

TABLE 1: Material parameters of Nitinol.

Parameters	Values
Austenite elasticity, $E_A$ (MPa)	51700
Austenite Poisson's ratio, $\nu_A$	0.3
Martensite elasticity, $E_M$ (MPa)	47800
Martensite Poisson's ratio, $\nu_M$	0.3
Transformation strain, $\epsilon^L$	0.063
Start of transformation loading, $\sigma_L^S$ (MPa)	600
End of transformation loading, $\sigma_L^E$ (MPa)	670

that high-magnitude strains are always located in the strut's inner fillet tensile side and its vicinity. The strain distributions across the MERD middle segments are negligible. The peak strain located in the "flow-prone" region achieves a value of 6.1%, which is much lower than the critical strain threshold

for Nitinol of 12% [10]. Therefore, no issues of cracks or fractures in the structure are likely in the shape-setting step.

**2.3. Migrating.** MERD struts may seriously contract or collapse due to contact-dominant bending in the tortuous blood vessel. The wall apposition performance is a significant criterion to evaluate the device effectiveness during migration and embolus capture. Figure 8 illustrates a significant cross-profile contraction of the middle segment struts while they are being pulled along a curved path.

The formula for the cross-profile reduction ratio is calculated and proposed as below:

$$R_1 = \frac{(D - L_1)}{D}, \quad (1)$$

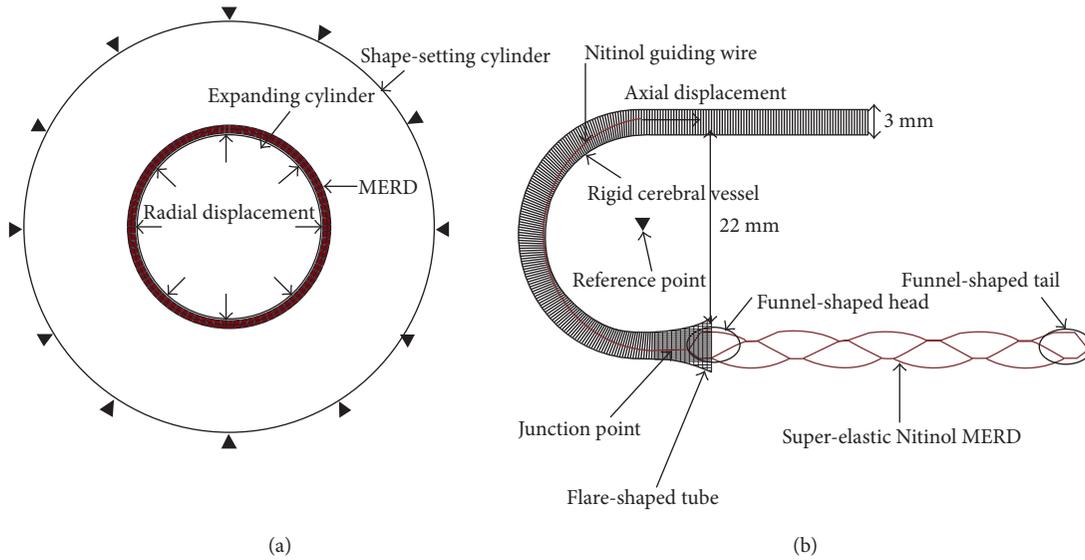


FIGURE 4: Schematic points of assembly boundary conditions. (a) Expansion for shape setting. (b) Pulling and migrating inside the tortuous cerebral vessel.

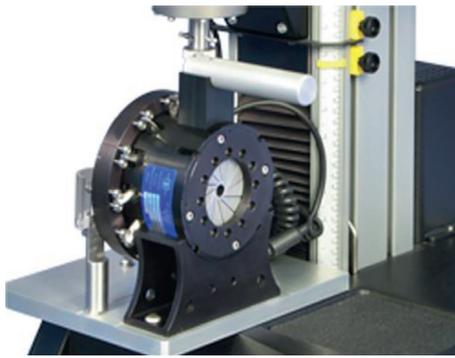


FIGURE 5: The electrical force sensor (RX500, Machine Solutions Inc.).

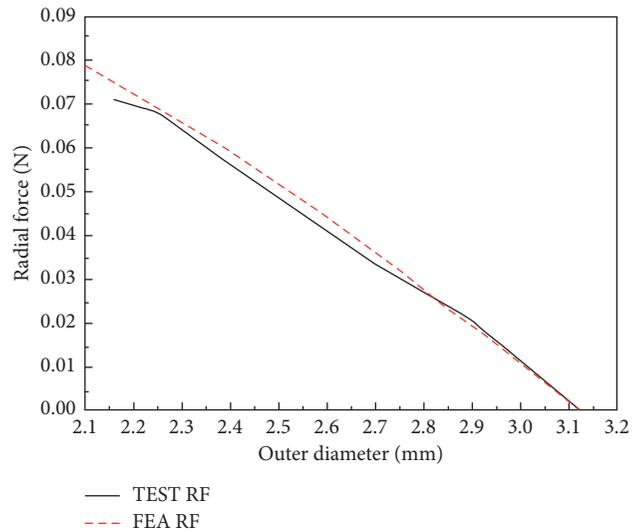


FIGURE 6: Experimental validation of FEA.

where  $D$  is the artery cross-profile diameter,  $L_1$  is the radial contraction lengths, and  $R_1$  is the cross-profile reduction ratios. As a result,  $D$  obtains a value of 3.0 mm,  $L_1$  0.8 mm, and  $R_1$  73.3%. The strain field distribution contour plots of the deformed MERD are displayed in Figure 9.

Figure 9 indicates that high-magnitude MPS is always located in the shoulder of the constrained strut's apex. In the migration step, a structure-induced resultant force is assumed to be applied to stretch and straighten the apex struts. The peak strain located on the 'funnel'-shaped head reaches a value of 7.0%. Compared to a Nitinol ultimate tensile strength (UTS) critical threshold of 12% [10], there is no risk of crack or fracture failure during the migration step. All other areas appear to have relatively low strain. They are also confirmed to stay in the safe domain of Nitinol alloy. It has been speculated that the former highest peak MPS appears on the basis of unsmooth geometry [17] and a tiny connecting section in the microstructure.

### 3. Discussion

This paper explores the biomechanical concept of the wall apposition performance for tubular MERD migration simulation. This behavior is regarded as reflecting the safety of pulling migration as well as the effects of embolus removal. Structure designers can simulate and analyze the virtual prototypes. Various topology and dimension optimization methods could also be used to provide closer contact with the tortuous vessel. As a vital mechanical index, the numerical wall apposition performance could offer surgeons a guideline for a reasonable MERD choice.

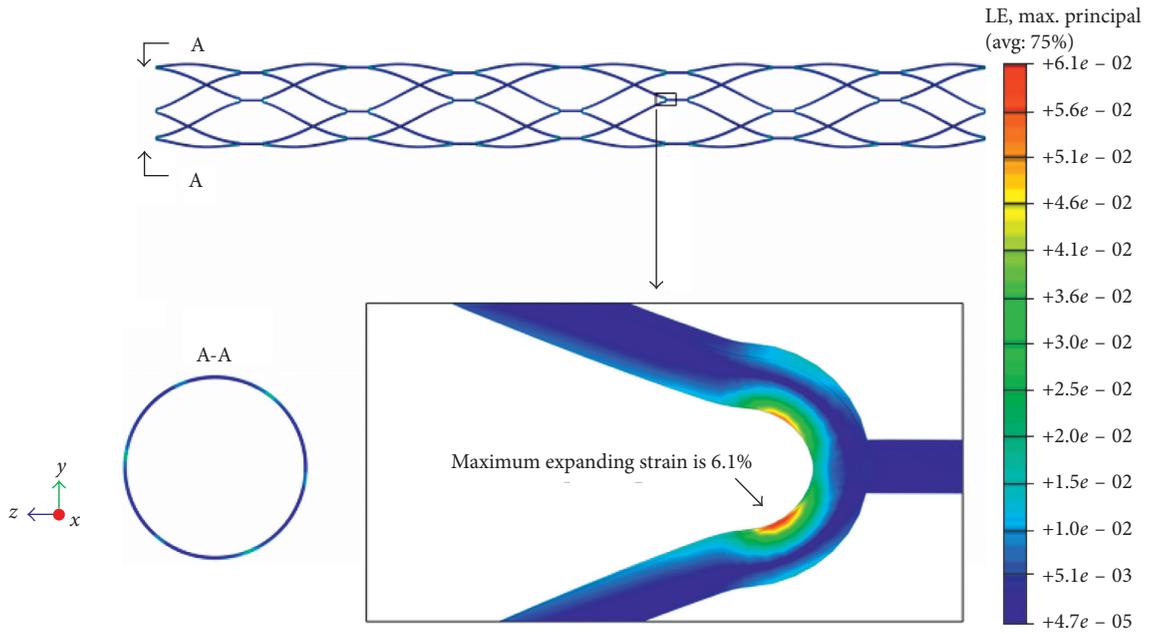


FIGURE 7: Maximum principal strain contour plots of MERD expansion: global and detailed schematic views.

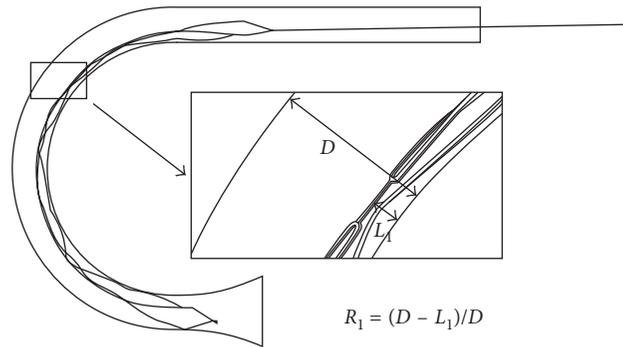


FIGURE 8: Wall apposition performance of a MERD segment in parametrical schematic.

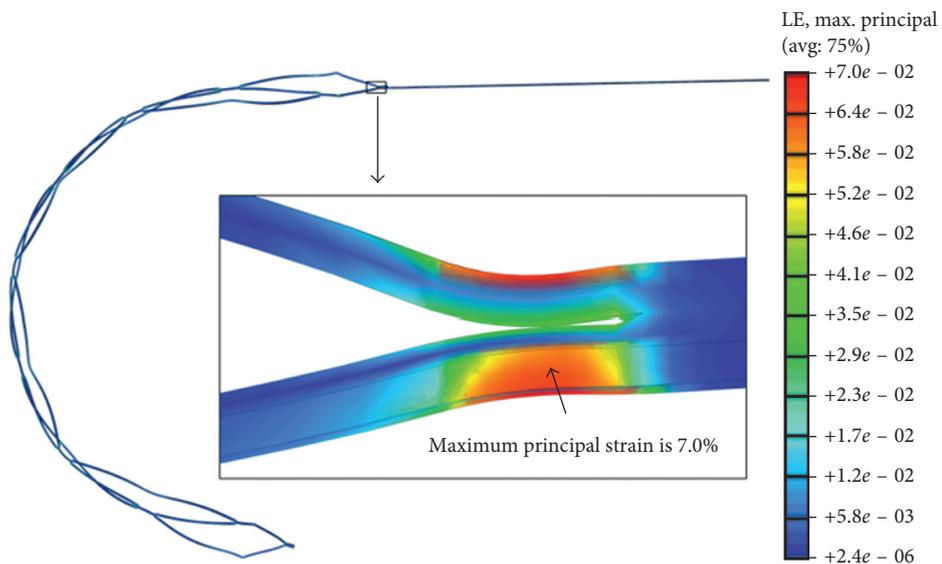


FIGURE 9: MPS field contour plots of MERD migration along a tortuous artery: global and detailed schematic views.

The contraction or collapse behavior is predicted to be induced by the bend resistance deficiency of the tubular axial structure. The key factors of the desired mechanical performance include the strut width, strut thickness, wave axial length, unit cell number, and connecting-bridge topology. The wall apposition numerical analysis of the MERD has been displayed in the absence of an arterial constitutive model and anatomical geometry. The simulation process of thrombus capture is also not presented. The simulation results can be applied to provide a reference for the design of stent retriever mechanical thrombectomy devices which achieve higher recanalization rates than nonstent retriever devices in acute ischemic stroke with improved clinical and radiographic outcomes and safety [11].

Based on the analysis above, a number of recommendations are suggested for engineers for optimization of MERD structure. For example, a shorter independent support unit cell could undergo and resist higher contact-induced axial bending loading. Therefore, an approach of axial length shortening can be utilized to improve the structure performance and minimize cross-sectional collapse effects. The geometry of the former highest peak MPS region should be optimized. To reduce the strut's apex strain, the strut's external side (the so-called "shoulder" region) and internal side (close to the fillet curve region) should be smooth and continuous. The radius of curvature of the fillet region should also be adjusted. Above analyses will provide valuable information about the MERD design and struts dimensions that can be optimized in order to maximize the effectiveness of the MERD during the shape-setting and migrating process.

In future work, a particular cerebral vessel wall will be three-dimensionally reconstructed and modeled on the basis of patients' MRI images. The cerebral vessel wall and emboli will be modeled as homogeneous isotropic hyperelastic materials. Two types of MERDs will be generated and simulated for the migration and embolus removal in the reconstructed cerebral vessel. During interventional mechanical thrombectomy, the management of the delivered system deployment and embolus capture is considered crucial and difficult to optimize. The method provided in the paper is suggested as an operational guideline for modeling accurate MERD insertion during clinical use. Based on the demonstrated axial shrinkage during MERD self-expansion, physicians can deliver more confidently the embolus device system to the correct position in a diseased cerebral vessel. In this way, the MERD will self-expand and capture plaque emboli more reliably. It will be better if some experiments are adopted to verify the results in the future.

#### 4. Conclusion

Our study suggests that the independent support unit cell should be reduced to resist higher contact-induced axial bending loading. An approach of axial length shortening can be utilized to improve the structure performance and minimize cross-sectional collapse effects. The geometry of the former highest peak MPS region should be optimized. In addition, these findings favor the use of verifying the

performance of the delivery system insertion and the effectiveness of embolus retrieval [13].

#### Conflicts of Interest

The authors declare no conflicts of interest.

#### Authors' Contributions

Xuelian Gu and Yongxiang Qi contributed equally to this work.

#### Acknowledgments

Xuelian Gu was partially supported by the National Natural Science Foundation of China (no. 11502146). Arthur Erdman was partially supported by the National Science Foundation (IIS-1251069) and the National Institutes of Health (1R01EB018205-01).

#### References

- [1] M. Desai, *Development of Asymmetric Stent Designs and Validation through Finite Element Analysis Simulations*, State University of New York at Buffalo, New York, NY, USA, 2006.
- [2] H. J. Shi, Y. H. Huang, T. Shen, and Q. Xu, "Percutaneous mechanical thrombectomy combined with catheter-directed thrombolysis in the treatment of symptomatic lower extremity deep venous thrombosis," *European Journal of Radiology*, vol. 71, no. 2, pp. 350–355, 2009.
- [3] Ö. Krischek, E. Miloslavski, S. Fischer, and S. Shrivastava, "A comparison of functional and physical properties of self-expanding intracranial stents [Neuroform3, Wingspan, Solitaire, Leo (+), Enterprise]," *Minimally Invasive Neurosurgery*, vol. 54, no. 1, pp. 21–28, 2011.
- [4] T. Kahles, C. Garcia-Esperon, S. Zeller et al., "Mechanical thrombectomy using the new ERIC retrieval device is feasible, efficient, and safe in acute ischemic stroke: a Swiss stroke center experience," *American Journal of Neuroradiology*, vol. 37, no. 1, pp. 114–119, 2016.
- [5] M. Goyal, A. M. Demchuk, B. K. Menon et al., "Randomized assessment of rapid endovascular treatment of ischemic stroke," *New England Journal of Medicine*, vol. 372, no. 11, pp. 1019–1030, 2015.
- [6] Z. S. Liu, L. J. Zhou, Y. Sun et al., "Thrombectomy using "clamping embolus with semi-retrieval" technique in acute ischemic stroke," *Journal of Stroke and Cerebrovascular Diseases*, vol. 27, no. 3, p. 733, 2018.
- [7] A. M. Spiotta, M. I. Chaudry, F. K. Hui et al., "Evolution of thrombectomy approaches and devices for acute stroke: a technical review," *Journal of NeuroInterventional Surgery*, vol. 7, no. 1, pp. 2–7, 2015.
- [8] J. Madjidyar, J. Hermes, S. Freitagwolf et al., "Stent-thrombus interaction and the influence of aspiration on mechanical thrombectomy: evaluation of different stent retrievers in a circulation model," *Neuroradiology*, vol. 57, no. 8, pp. 791–797, 2015.
- [9] D. Fiorella, "The evolution of stenting and stent-retrieval for the treatment of acute ischemic stroke," *Cardiovascular Engineering and Technology*, vol. 4, no. 4, pp. 352–356, 2013.
- [10] C. Kleinstreuer, Z. Li, C. A. Basciano, and S. Seelecke, "Computational mechanics of Nitinol stent grafts," *Journal of Biomechanics*, vol. 41, no. 11, pp. 2370–2378, 2008.

- [11] K. A. Hentschel, B. Daou, N. Chalouhi et al., "Comparison of non-stent retriever and stent retriever mechanical thrombectomy devices for the endovascular treatment of acute ischemic stroke," *Journal of Neurosurgery*, vol. 126, no. 4, pp. 1123–1130, 2017.
- [12] M. Azaouzi, A. Makradi, and S. Belouettar, "Deployment of a self-expanding stent inside an artery: a finite element analysis," *Materials and Design*, vol. 41, pp. 410–420, 2012.
- [13] X. Gu, Y. Qi, A. Erdman et al., "The role of simulation in the design of a semi-enclosed tubular embolus retrieval," *Journal of Medical Devices*, vol. 11, pp. 210011–210017, 2017.
- [14] W. Wu, D. Z. Yang, M. Qi, and W. Q. Wang, "An FEA method to study flexibility of expanded coronary stents," *Journal of Materials Processing Technology*, vol. 184, no. 1, pp. 447–450, 2007.
- [15] J. A. Grogan, S. B. Leen, and P. E. McHugh, "Comparing coronary stent material performance on a common geometric platform through simulated bench testing," *Journal of the Mechanical Behavior of Biomedical Material*, vol. 12, pp. 129–138, 2012.
- [16] M. Azaouzi, A. Makradi, and S. Belouettar, "Numerical investigations of the structural behavior of a balloon expandable stent design using finite element method," *Computational Materials Science*, vol. 72, pp. 54–61, 2013.
- [17] E. Masoumi Khalil Abad, D. Pasini, and R. Cecere, "Shape optimization of stress concentration-free lattice for self-expandable Nitinol stent-grafts," *Journal of Biomechanics*, vol. 45, no. 6, pp. 1028–1035, 2012.

## Research Article

# A Filtering Method for Identification of Significant Target mRNAs of Coexpressed and Differentially Expressed MicroRNA Clusters

Su Yeon Lee,<sup>1</sup> Soo-Yong Shin,<sup>2</sup> Young Jo Yoon,<sup>3</sup> and Yu Rang Park<sup>4</sup> 

<sup>1</sup>Bioinformatics Team, Samsung SDS, Seoul, Republic of Korea

<sup>2</sup>Department of Digital Health, SAIHST, Sungkyunkwan University, Seoul, Republic of Korea

<sup>3</sup>Office of Clinical Research Information, Asan Medical Center, Seoul, Republic of Korea

<sup>4</sup>Department of Biomedical Systems Informatics, Yonsei University College of Medicine, Seoul, Republic of Korea

Correspondence should be addressed to Yu Rang Park; [yurang.park@amc.seoul.kr](mailto:yurang.park@amc.seoul.kr)

Received 2 October 2017; Revised 8 February 2018; Accepted 16 July 2018; Published 12 September 2018

Academic Editor: Yong Xia

Copyright © 2018 Su Yeon Lee 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.

MicroRNA (miRNA) binding is primarily based on sequence, but structure-specific binding is also possible. Various prediction algorithms have been developed for predicting miRNA target genes; the results, however, have relatively high levels of false positives, and the degree of overlap between predicted targets from different methods is poor or null. We devised a new method for identifying significant miRNA target genes from an extensive list of predicted miRNA target gene relationships using hypergeometric distributions. We evaluated our method in statistical and semantic aspects using a common miRNA cluster from six solid tumors. Our method provides statistically and semantically significant miRNA target genes. Complementing target prediction algorithms with our proposed method may have a significant synergistic effect in finding and evaluating functional annotation and enrichment analysis for miRNA.

## 1. Introduction

MicroRNAs (miRNAs) are a class of small RNAs that regulate gene expression at the transcript level, protein level, or both [1–4]. miRNAs modulate gene activity and are aberrantly expressed in most types of cancers [5]. Due to their small size and stability, miRNAs can also be measured in biologic fluids such as plasma and serum and can serve as circulating biomarkers [6–9]. In spite of the continuous attempts to identify miRNAs and to elucidate their basic mechanisms of action, little is understood about their biological functions.

Because of the regulatory role of miRNAs and lack of direct functional annotation to miRNAs, functional enrichment methods for miRNAs rely on their target gene's functional annotations [10–12]. For instance, if the target genes of a specific miRNA are significantly enriched with a set of Gene Ontology (GO) terms, it is reasonable to infer

that the miRNA is also involved in the same GO annotations. Several studies on miRNAs have used “predicted target-genes’ functional annotation-based” miRNA function prediction strategy [13, 14]; these methods, however, are limited in that they do not consider the many-to-many-to-many tripartite network topology among miRNAs, target genes, and GO annotation [15–17]. In our previous work, we proposed three types of measures (miRNA-centric, target gene-centric, and target link-centric) and a novel index for calculating the functional enrichment of miRNA. Among the three measures, the miRNA-centric measurements showed the best performance [18]. We also found that the miRNA’s intrinsic properties of multiplicity and cooperability may be correctly modeled by combined hypergeometric distributions.

Most of the miRNA-to-mRNA target links are estimated by prediction algorithms. However, these algorithms generate a relatively high level of false positives [19], and the

degree of overlap between predicted targets from different methods is often poor or null [13, 20]. Studies in this field have developed multiple databases with enormous amount of miRNA-to-target mRNA relationships computed using diverse algorithms [21], whereas only a few experimentally validated targets are available [22, 23]. In light of this circumstance, there is an unmet need for a method for identifying a significant miRNA target from a copious amount of predicted miRNA-target mRNA pairs.

According to miRNA characteristics (multiplicity and cooperatively activities), we employed the hypergeometric distribution to identify significant miRNA target genes from the extensive list of miRNA target genes. We also evaluated the performance of our method in two aspects: statistical significance and functional enrichment.

## 2. Methods

**2.1. Computational Methods.** To find significant target mRNAs from the input miRNA cluster, we first searched for target mRNAs from all miRNA members within the input cluster from the miRNA target database. For each targeted mRNAs, we then calculated the numbers of miRNAs that have target relationships ( $p_i$ ) with the mRNA and those that do not ( $p_j$ ) using the two-by-two contingency table. We also calculated the numbers of miRNAs not in the input miRNA cluster by dividing those that have a target relationship ( $p_k$ ) with the mRNA and those that do not ( $p_l$ ), as shown in Table 1. Functional enrichment was tested from this contingency table using a hypergeometric distribution. The hypergeometric distribution applies to sampling without replacement from a finite population whose elements can be classified into two mutually exclusive categories: has/does not have a target relationship.

We then calculated the adjusted  $p$  values using the Bonferroni correction. Finally, for evaluating our methods, 10,000 simulated mRNA sets of the same size were also randomly sampled from the target mRNAs of the input miRNA cluster.

Using hypergeometric distribution, we assumed that the coordinated function among miRNAs within a cluster is valid when these miRNAs are regulated or annotated by common factors such as same target mRNA, Gene Ontology, or pathway.

**2.2. Data Set: miRNA Clusters.** We obtained an miRNA set created by Volinia et al. [24] that has differentially expressed sets of up- or downregulated miRNAs in six solid tumor samples. Among the miRNA clusters, we selected an miRNA cluster composed of 57 miRNAs by prediction analysis of microarray (PAM) in six solid tumor samples versus normal tissues. The complete list of 57 miRNAs is in Additional File 1.

**2.3. Creating Variations of the miRNA-mRNA Target Pair.** To build the miRNA-mRNA target pair, we chose three representative miRNA databases: TarBase (Data release 6.0,

TABLE 1:  $2 \times 2$  contingency table of miRNA frequency calculated for each target mRNA.

Target mRNA	Input miRNA cluster	
	In cluster	Not in cluster
Has a target relationship	$p_i$	$p_k$
Does not have a target relationship	$p_j$	$p_l$

February 28<sup>th</sup>, 2014) [23], MirTarBase (Data release 4.5, February 28<sup>th</sup>, 2014) [22], and mirDIP (Data release 1.0, February 12<sup>th</sup>, 2014) [14]. The TarBase and MirTarBase databases provide experimentally validated miRNA-target interaction data and evidence level (strong and less strong) of each interaction. The mirDIP database provides *in silico*-predicted miRNA-target interaction data from six established target prediction algorithms and 12 miRNA prediction databases. GO annotation of miRNA-target-mRNA was obtained from the Entrez Gene database. We excluded GO associations with ND (no biological data) and NR (not recorded) evidence code. Detailed processes are provided in Figure 1.

**2.4. Statistical and Semantic Evaluation Measurement.** To evaluate our methods, we compared the performance in terms of statistical significances between a significant 317 mRNA cluster and randomly simulated 10,000 clusters. Each randomly generated cluster had the same size as the significant mRNA set. GO functional enrichment analysis was performed for all mRNA sets using GO annotations retrieved from the NCBI Entrez Gene database. We filtered out 339 GO terms that were greater than 0.05. The resulting lists of 377 GO terms are shown in Additional File 2. To reduce the number of GO terms, enriched GO terms and  $p$  values were submitted to REduce and Visualize GO (REViGO).

We computed the average log  $p$  values of ranked GO term sets from functional enrichment analysis of the significant mRNA set and the randomly simulated 10,000 mRNA sets. These average log  $p$  values were then used for comparing the performance. Functional enrichment was performed using GO annotations of mRNA from NCBI Entrez gene [25]. Average log  $p$  values of ranked GO terms were based on the general assumption that highly significant GO terms are more desirable because it means the members of the cluster are highly correlated to each other. For semantic evaluation of the significant mRNA set, we used REViGO, which is a web-based system that summarizes a list of GO terms by finding a representative subset of the terms using the semantic similarity-based clustering algorithm [26].

## 3. Results

**3.1. Statistical Significances.** For evaluating our methods, we compare performance in terms of statistical significances between a significant 317 mRNA cluster and randomly

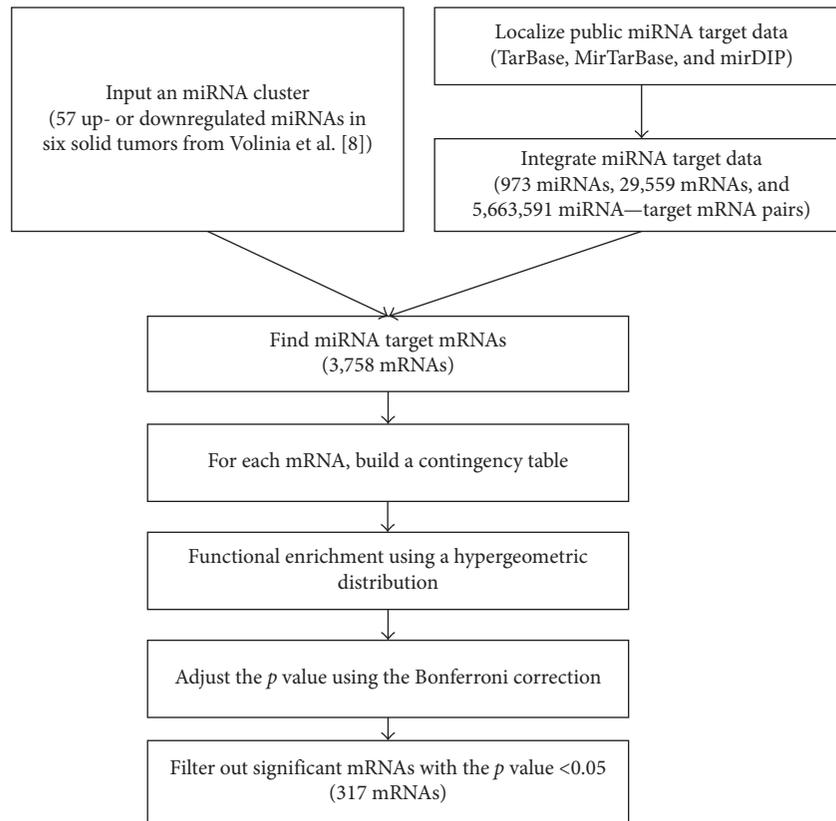


FIGURE 1: Flowchart of the computational method for identifying significant miRNA target genes.

simulated 10,000 clusters. Each randomly generated cluster has the same size with the significant mRNA cluster. All mRNA cluster performed GO functional enrichment analysis using GO annotation from the NCBI Entrez gene. Figure 2 shows the distributions of average  $\log p$  values for the rank of GO terms which belong to the biological process category. The significant mRNA set is shown as the red dotted graph. Randomly simulated 10,000 clusters are shown as box plots. The significant mRNA set showed a higher average  $\log p$  value than random clusters did, which indicated that the members of the cluster highly correlated and meaningfully composed.

**3.2. Gene Ontology Analysis of Significant miRNA Target Genes.** Using the UniProt database as background and the default semantic measure (SimRel), our analysis clearly showed that biological processes associated with cancer metabolism, regulation of cell death and apoptotic process, and negative regulation of autophagy were significantly overrepresented.

Figure 3 shows the REViGO scatter plot represented in a two-dimensional space derived by applying multidimensional scaling to a matrix of GO terms semantic similarities. The resulting lists of 339 GO terms along with their  $p$  values were further summarized by the REViGO reduction analysis tool that condenses the GO description by removing redundant terms. The remaining

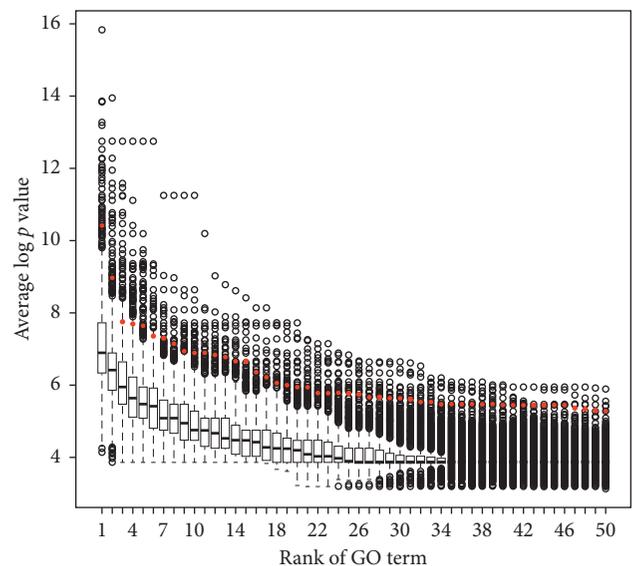


FIGURE 2: Evaluation of statistical significance across thresholds. The significant mRNA set and randomly simulated 10,000 clusters are shown as red dotted graphs and box plots, respectively.

terms after the redundancy reduction were plotted in a two-dimensional space. Bubble color indicates the  $p$  value (legend in the upper right-hand corner): the two ends of the colors are red and blue, which represent lower and higher  $p$  values, respectively. Size indicates the

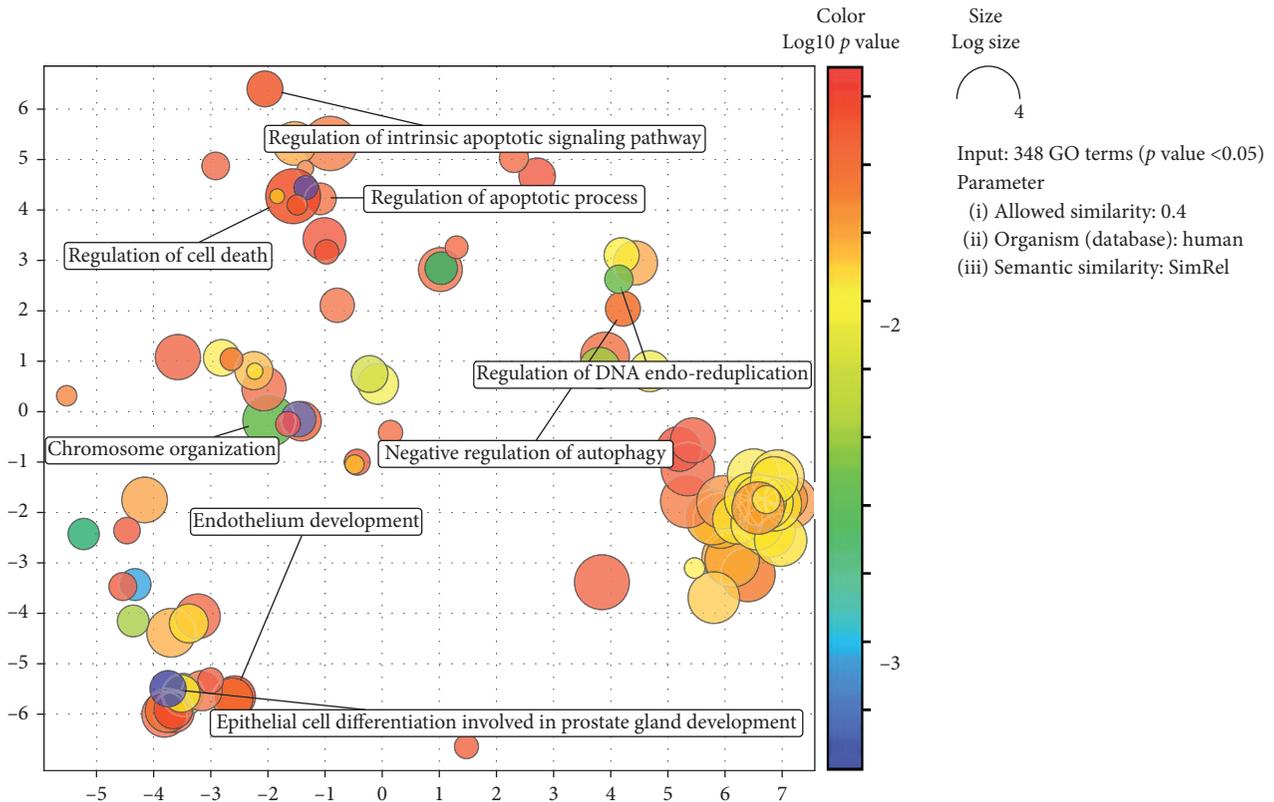


FIGURE 3: REViGO scatter plot for the significant mRNA set.

relative frequency of the GO term in the underlying reference UniProt databases (more general terms are represented by larger size bubbles).

#### 4. Discussion

Functional enrichment studies for miRNA expression are performed in three steps: (1) selecting differentially expressed miRNAs, (2) finding their target mRNA, and (3) carrying out analysis of mRNA set overrepresentation [27]. Functional enrichment studies for miRNAs are mostly based on the annotation of target mRNA; however, due to improvements in the miRNA target prediction algorithms, a large number of target mRNAs are predicted. Considering this, filtering out significant mRNAs using a stable statistical method is of great importance. In this study, we proposed a method for identifying the significant miRNA target mRNA from the miRNA cluster. The proposed method was verified by functional enrichment analysis of differentially expressed or coexpressed miRNA clusters.

Inaccurate functional enrichment methods are a hindrance in increasing clinical utility for miRNAs, such as miRNA-based biomarkers or predictors [28, 29]. Several tools have been recently established for direct prediction of miRNA functions [10, 30]; however, these methods do not consider the regulatory or indirect functions of miRNAs, such as regulation or inhibition of target genes [31]. The

intrinsic properties of multiplicity and cooperative activities of miRNAs should be considered while annotating the miRNA function. miRGator v3.0 is a tool created considering these characteristics and allows the user to manually select miRNAs and target mRNAs [32]. However, such tools are only useful when the number of miRNA and mRNA pairs is small.

The limitation of the proposed method is that the hypergeometric distribution has a significant effect when members belonging to an miRNA cluster are regulated by common factors such as the target mRNA, GO, and pathway. The proposed method constructs a target mRNA set with statistical significance by receiving miRNA clusters with similar expression characteristics. The assumption of hypergeometric is well suited to this problem because the cluster-received input already has similar characteristics.

The miRNA target prediction algorithms were modified to generate more accurate results based on the expanding understanding of the molecular mechanism of miRNA regulation. Nevertheless, identifying significant target mRNAs from the numerous, uncurated miRNA target links remain as a problem. Our method is based on computationally identifying statistically significant mRNAs using predicted or experimentally validated target relationships. Complementing target prediction algorithms with our proposed method may have significant synergistic effects in finding and evaluating functional annotation and enrichment analysis for miRNA.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Authors' Contributions

Su Yeon Lee and Soo-Yong Shin have contributed equally to this work.

## Acknowledgments

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2017R1D1A1B03035762).

## Supplementary Materials

*Supplementary 1.* Additional File 1: the complete list of 57 miRNAs.

*Supplementary 2.* Additional File 2: the resulting lists of 377 GO terms.

## References

- [1] D. P. Bartel, "MicroRNAs: target recognition and regulatory functions," *Cell*, vol. 136, no. 2, pp. 215–233, 2009.
- [2] P. T. Nelson, D. A. Baldwin, L. M. Scearce, J. C. Oberholtzer, J. W. Tobias, and Z. Mourelatos, "Microarray-based, high-throughput gene expression profiling of microRNAs," *Nature Methods*, vol. 1, no. 2, pp. 155–161, 2004.
- [3] E. C. Lai, "microRNAs: runts of the genome assert themselves," *Current Biology*, vol. 13, no. 23, pp. R925–R936, 2003.
- [4] K. Sun and E. C. Lai, "Adult-specific functions of animal microRNAs," *Nature Reviews Genetics*, vol. 14, no. 8, pp. 535–548, 2013.
- [5] M. V. Iorio and C. M. Croce, "MicroRNA dysregulation in cancer: diagnostics, monitoring and therapeutics. A comprehensive review," *Embo Molecular Medicine*, vol. 9, no. 6, p. 852, 2017.
- [6] X. Chen, Y. Ba, L. Ma et al., "Characterization of microRNAs in serum: a novel class of biomarkers for diagnosis of cancer and other diseases," *Cell Research*, vol. 18, no. 10, pp. 997–1006, 2008.
- [7] P. S. Mitchell, R. K. Parkin, E. M. Kroh et al., "Circulating microRNAs as stable blood-based markers for cancer detection," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 30, pp. 10513–10518, 2008.
- [8] G. Sozzi, M. Boeri, M. Rossi et al., "Clinical utility of a plasma-based miRNA signature classifier within computed tomography lung cancer screening: a correlative MILD trial study," *Journal of Clinical Oncology*, vol. 32, no. 8, pp. 768–773, 2014.
- [9] R. Belzeaux, R. X. Lin, and G. Turecki, "Potential use of microRNA for monitoring therapeutic response to antidepressants," *CNS Drugs*, vol. 31, no. 4, pp. 253–262, 2017.
- [10] I. Ulitsky, L. C. Laurent, and R. Shamir, "Towards computational prediction of microRNA function and activity," *Nucleic Acids Research*, vol. 38, no. 15, p. e160, 2010.
- [11] Y. Q. Wu, D.-J. Chen, H.-B. He et al., "Pseudorabies virus infected porcine epithelial cell line generates a diverse set of host microRNAs and a special cluster of viral microRNAs," *PLoS One*, vol. 7, no. 1, Article ID e30988, 2012.
- [12] Y. Xiao, C. Xu, J. Guan et al., "Discovering dysfunction of multiple microRNAs cooperation in disease by a conserved microRNA co-expression network," *PLoS One*, vol. 7, no. 2, Article ID e32201, 2012.
- [13] S. Ekimler and K. Sahin, "Computational methods for microRNA target prediction," *Genes*, vol. 5, no. 3, pp. 671–683, 2014.
- [14] E. A. Shirdel, W. Xie, T. W. Mak, and I. Jurisica, "NAVIGating the micronome—using multiple microRNA prediction databases to identify signalling pathway-associated microRNAs," *PLoS One*, vol. 6, no. 2, Article ID e17429, 2011.
- [15] D. Gaidatzis, E. van Nimwegen, J. Hausser, and M. Zavolan, "Inference of miRNA targets using evolutionary conservation and pathway analysis," *BMC Bioinformatics*, vol. 8, no. 1, p. 69, 2007.
- [16] J. Xu and C. Wong, "A computational screen for mouse signaling pathways targeted by microRNA clusters," *RNA*, vol. 14, no. 7, pp. 1276–1283, 2008.
- [17] Y. Gusev, "Computational methods for analysis of cellular functions and pathways collectively targeted by differentially expressed microRNA," *Methods*, vol. 44, no. 1, pp. 61–72, 2008.
- [18] S. Y. Lee, K. A. Sohn, and J. H. Kim, "MicroRNA-centric measurement improves functional enrichment analysis of co-expressed and differentially expressed microRNA clusters," *BMC Genomics*, vol. 13, no. 7, p. S17, 2012.
- [19] L. S. Hon and Z. Zhang, "The roles of binding site arrangement and combinatorial targeting in microRNA repression of gene expression," *Genome Biology*, vol. 8, no. 8, p. R166, 2007.
- [20] P. Sethupathy, M. Megraw, and A. G. Hatzigeorgiou, "A guide through present computational approaches for the identification of mammalian microRNA targets," *Nature Methods*, vol. 3, no. 11, pp. 881–886, 2006.
- [21] W. Ritchie and J. E. Rasko, "Refining microRNA target predictions: sorting the wheat from the chaff," *Biochemical and Biophysical Research Communications*, vol. 445, no. 4, pp. 780–784, 2014.
- [22] S. D. Hsu, F.-M. Lin, W.-Y. Wu et al., "miRTarBase: a database curates experimentally validated microRNA-target interactions," *Nucleic Acids Research*, vol. 39, no. 1, pp. D163–D169, 2011.
- [23] T. Vergoulis, I. S. Vlachos, P. Alexiou et al., "TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support," *Nucleic Acids Research*, vol. 40, no. D1, pp. D222–D229, 2012.
- [24] S. Volinia, G. A. Calin, C.-G. Liu et al., "A microRNA expression signature of human solid tumors defines cancer gene targets," *Proceedings of the National Academy of Sciences*, vol. 103, no. 7, pp. 2257–2261, 2006.
- [25] NCBI Resource Coordinators, "Database resources of the National Center for Biotechnology Information," *Nucleic Acids Research*, vol. 43, no. D1, 2014.
- [26] F. Supek, M. Bošnjak, N. Škunca, and T. Šmuc, "REVIGO summarizes and visualizes long lists of gene ontology terms," *PLoS One*, vol. 6, no. 7, Article ID e21800, 2011.
- [27] F. Garcia-Garcia, J. Panadero, J. Dopazo, and D. Montaner, "Integrated gene set analysis for microRNA studies," *Bioinformatics*, vol. 32, no. 18, pp. 2809–2816, 2016.
- [28] T. Bleazard, J. A. Lamb, and S. Griffiths-Jones, "Bias in microRNA functional enrichment analysis," *Bioinformatics*, vol. 31, no. 10, pp. 1592–1598, 2015.

- [29] W. Ritchie, S. Flamant, and J. E. Rasko, "Predicting microRNA targets and functions: traps for the unwary," *Nature Methods*, vol. 6, no. 6, pp. 397-398, 2009.
- [30] M. Lu, B. Shi, J. Wang, Q. Cao, and Q. Cui, "TAM: a method for enrichment and depletion analysis of a microRNA category in a list of microRNAs," *BMC Bioinformatics*, vol. 11, no. 1, p. 419, 2010.
- [31] L. P. Lim, N. C. Lau, P. Garrett-Engele et al., "Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs," *Nature*, vol. 433, no. 7027, pp. 769-773, 2005.
- [32] S. Cho, I. Jang, Y. Jun et al., "miRGator v3.0: a microRNA portal for deep sequencing, expression profiling and mRNA targeting," *Nucleic Acids Research*, vol. 41, no. D1, pp. D252-D257, 2013.

## Research Article

# Comparative Study on Automated Cell Nuclei Segmentation Methods for Cytology Pleural Effusion Images

**Khin Yadanar Win** <sup>1</sup>, **Somsak Choomchuay**,<sup>1</sup> **Kazuhiko Hamamoto**,<sup>2</sup>  
**and Manasanan Raveesunthornkiat** <sup>3</sup>

<sup>1</sup>Faculty of Engineering, King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand

<sup>2</sup>School of Information and Telecommunication Engineering, Tokai University, Tokyo, Japan

<sup>3</sup>Department of Pathology, Faculty of Medicine, Srinakharinwirot University, Nakhon Nayok, Thailand

Correspondence should be addressed to Khin Yadanar Win; [kyadanarw@gmail.com](mailto:kyadanarw@gmail.com)

Received 17 November 2017; Accepted 18 July 2018; Published 12 September 2018

Academic Editor: Weidong Cai

Copyright © 2018 Khin Yadanar Win 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.

Automated cell nuclei segmentation is the most crucial step toward the implementation of a computer-aided diagnosis system for cancer cells. Studies on the automated analysis of cytology pleural effusion images are few because of the lack of reliable cell nuclei segmentation methods. Therefore, this paper presents a comparative study of twelve nuclei segmentation methods for cytology pleural effusion images. Each method involves three main steps: preprocessing, segmentation, and postprocessing. The preprocessing and segmentation stages help enhancing the image quality and extracting the nuclei regions from the rest of the image, respectively. The postprocessing stage helps in refining the segmented nuclei and removing false findings. The segmentation methods are quantitatively evaluated for 35 cytology images of pleural effusion by computing five performance metrics. The evaluation results show that the segmentation performances of the Otsu, k-means, mean shift, Chan-Vese, and graph cut methods are 94, 94, 95, 94, and 93%, respectively, with high abnormal nuclei detection rates. The average computational times per image are 1.08, 36.62, 50.18, 330, and 44.03 seconds, respectively. The findings of this study will be useful for current and potential future studies on cytology images of pleural effusion.

## 1. Introduction

Globally, cancer is one of the deadliest diseases with high morbidity and mortality. In 2015, approximately 14 million new cases were diagnosed, and over 8 million deaths were estimated worldwide [1]. Ferlay et al. [2] estimated that the death toll due to cancer is set to rise dramatically by approximately 70% in the coming decades. Fortunately, the mortality and morbidity associated with cancer can be reduced, with a high potential for cure if cancer is diagnosed and treated at an early stage.

When cancer grows or flows in the pleura cavity between the lungs and the chest wall, it causes a malignant pleural effusion, which is the excessive collection of pleural fluid, as shown in Figure 1 [3].

Fifty percent of cancer patients have a high possibility of developing malignant pleural effusion. Both primary and metastasis cancers can be diagnosed from pleural effusion

[4]. The most frequently occurring primary cancer in pleural effusion is mesothelioma. The most common types of metastasis cancers are the cancers of the breasts, lungs, ovaries, and blood, including other unknown primary sites.

Pleural effusion can be detected using several imaging approaches such as X-ray, ultrasound, computed tomography (CT), and magnetic resonance imaging (MRI), including other tests such as urine and blood tests. For the assessment of malignancy, a cytological examination is performed by pathologists because it is simple, cheap, less invasive, and highly effective.

The cytological exam is a manual procedure wherein cytologists or experts visually investigate every single cell on the cytology glass slides using a microscopic camera, identify any abnormality in a cell, and finally make a decision. However, the procedure is time-consuming and requires good skill; moreover, it is tedious and prone to inter- and intraobserver variations. In addition, the diagnosis accuracy

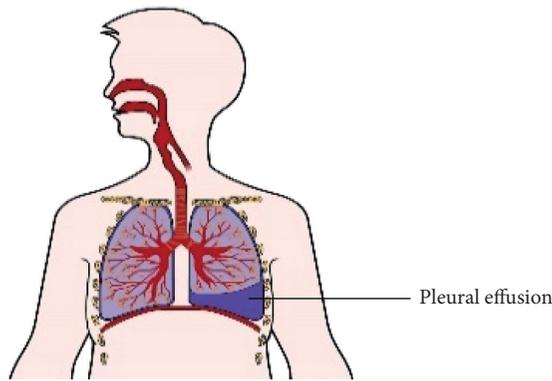


FIGURE 1: Excessive amount of pleural effusion.

strongly depends on the attention and expertise of cytologists. These factors have inspired us to implement a computer-aided diagnosis (CAD) system. The CAD system can help relieve the workload on cytologists, accelerate the diagnosis procedure, eliminate the inter- and intraobserver variations in the diagnosis, and describe the quantitative results, thus complementing and enhancing the opinions of the cytologists. To implement a CAD system for cancer cells, cell nuclei segmentation is a prerequisite because cancer cells are largely observed on the basis of the morphological changes in the cell nuclei. Therefore, it is crucial to select an accurate and effective cell nuclei segmentation method that can help precisely delineate the nuclei contours.

## 2. Literature Review

Several promising nuclei segmentation approaches have been proposed for different types of microscopic images, for example, cervical and breast cells. Cell nuclei segmentation methods for breast cell images [5–10] and cervical cell images [11–15] have been reported. Moreover, methods have been proposed to segment cell nuclei on prostate cancer cell images [16, 17] and microscopic blood cell images [18, 19]. Significant efforts have been made to segment the cell nuclei from other types of images such as bone marrow images, lung sputum cells images, brain cell images, liver and thyroid images, bladder and skin tissue images, and brain glioma images in various studies [20–24], respectively.

Although many robust and effective methods for nuclei segmentation have been reported [5–24], they are designed to address specific types of cell images such as those of the breast, cervical, blood, and prostate. In different types of cell images, the structure of the cells and their corresponding gray-level distributions vary significantly. Therefore, the aforementioned methods cannot be directly employed to segment all types of images on the basis of the diversity of the images.

Only a few studies have detected malignant cells from cytology pleural effusion images. Zhang et al. [25] reported a method to detect malignant cells from pleural effusion images using integrated fuzzy edge detection and Otsu's method. Chen et al. [26] presented a method on the basis of the wavelet and morphology transform to detect malignant cells from pleural effusion images.

However, a preprocessing stage for removing noise and enhancing contrast is not considered in these methods, thus reducing the accuracy of the detection system. In addition, the methods are not focused on the segmentation process and there is a lack of quantitative evaluation of the proposed methods. To address the shortcomings, we recently reported two cell nuclei segmentation methods [27, 28]. However, nuclei segmentation in cytology pleural effusion images is still a challenging task because of the high cell population, their varieties, overlapping cells, and poor cell contrast. There is scope to further improve the cell segmentation method for cytology pleural effusion images. Thus, more observations are required to implement and determine the most feasible segmentation method.

In the cytology and histology image analysis, nuclei segmentation often revolves around thresholding techniques, clustering techniques, and active contour techniques. Thresholding techniques are quite simple ones in nuclei segmentation. Every pixel in the image is determined into nuclei or background depending on the image intensity. In spite of providing the effective segmentation performance for the images with uniform background, they are sensitive to noises, uneven background, and intensity heterogeneity inside the images [17, 25, 29, 30, 31]. Clustering techniques are unsupervised methods that attempt to group the pixels having similar features into different objects without prelabeling the samples [10, 14, 19, 20]. Active contour techniques are based on moving the deformable splines inside the images to find nuclei contours using the gradient information [6, 16, 32, 33]. Graph cut methods recently have a great interest in nuclei segmentation and yield the good segmentation performance [9, 11]. Other popular techniques are variants of watershed method and concavity analysis method which are frequently used for isolating the overlapping nuclei that is critical issue of cytology and histology image segmentation. The above stated techniques are widely utilized in either individual or integrating together in many microscopy image analyses, especially in cervical cytology and breast histology image segmentation. According to their simplicity, reproducibility and affordable cost of processing methods are (still) considered in our new designs of cytology pleural effusion analysis. We emphasize on simple and classical image segmentation methods, that is, thresholding, clustering, active contour, and graph cut methods, to extract the nuclei region from the background in cytological pleural effusion images.

In this paper, we experimentally employed twelve cell nuclei segmentation methods individually from four techniques: (i) thresholding techniques, (ii) clustering techniques, (iii) graph cut techniques, and (iv) active contours. Each method involves three main stages: preprocessing, segmentation, and postprocessing. In the preprocessing stage, contrast-limited adaptive histogram equalization (CLAHE) and median filtering are used to enhance the image quality and remove the small noises. In the segmentation stage, we proposed twelve segmentation methods, including (1) Otsu's method, (2) Isodata thresholding method, (3) maximum entropy thresholding method, (4) cross entropy thresholding, (5) minimum error thresholding, (6) fuzzy entropy thresholding method, (7) adaptive thresholding method, (8) k-means

clustering, (9) fuzzy c-means clustering, (10) mean shift clustering, (11) Chan–Vese level set, and (12) graph cut methods. Finally, the boundaries of the segmented cell nuclei are refined, and false findings are eliminated using morphological methods. The methods are quantitatively evaluated in terms of five performance metrics, and their accuracies are later compared. Finally, a discussion along with the advantages and disadvantages of the methods is presented.

The examined methods aim to extract the nuclei regions from the surrounding objects and background in the images. These methods attempt to segment out all possible nuclei regions regardless of single laying or overlapping. They are not capable of separating the overlapping cells. It should also be noted that the performance metrics measured here are aimed to compute the correctness of the segmented pixels by matching with the pixels in the hand-drawn ground truth image. When computing the nuclei detection rate, each connected region from the segmentation results is considered as one nucleus regardless of the number of nucleus inside the region. The overlapping issue is not taken into account in the evaluation processes of segmentation accuracy and nuclei detection rate. Nevertheless, the undersegmentation errors of the overlapping cells can affect the final decision of CAD system. Thus, the isolation of overlapping nuclei and extraction of interregional walls will be remained as future study. On the bright side, the examined methods have potential to integrate with overlapped splitting methods to split the overlapped nuclei.

The rest of this paper is structured as follows. Section 3 presents the datasets and segmentation methods. Section 4 presents a benchmark of the experimental results, including parameters tuning and discussion. Finally, the conclusions of this study are presented in Section 5.

### 3. Materials and Methods

#### 3.1. Dataset Description and Ground Truth Segmentation.

We are not aware of any publicly available dataset for cytology pleural effusion images. Thus, we prepared a local dataset with the help of a hospital. The studied dataset is based on cytology glass slides of pleural effusion specimens obtained with the cooperation of experts from the Department of Pathology, Faculty of Medicine, Srinakharinwirot University, Thailand. The samples were taken from the pleural space using thoracentesis procedure, spread on the glass slides, and stained using classical Papanicolaou (Pap) staining method. The images were captured using an Olympic microscope mounted on the digital camera. The dataset comprises 35 images of cytology pleural effusion containing healthy cells, benign cells, and cancerous cells. The resolutions of the images were  $4050 \times 2050$  pixels, stored in 8-bit RGB space. Figure 2 shows the sample and component of the cytology pleural effusion image. To set the gold standard, the ground truth images were prepared with the help of experts from the hospital. First, computer vision researchers manually delineated the cell nuclei. The experts then verified and annotated the cell nuclei and pathology cells.

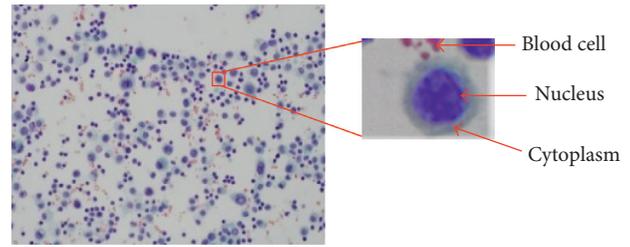


FIGURE 2: Pap smear images of pleural effusion and its components.

**3.2. Cell Nuclei Segmentation Framework.** Three main stages are considered to automatically segment the cell nuclei. Figure 3 shows the segmentation framework. In the next section, we introduce the details of each stage.

**3.3. Preprocessing Stage.** The preprocessing stage is an essential step in improving the quality of the image. First, to reduce the computational load, the original input image is resized to resolutions of  $1052 \times 1052$  pixels. The resized image is then converted into different color spaces using the segmentation methods. The cytology images might contain debris, noises, or stained artefacts because of the uneven illumination or dirt on the camera surface resulting from the image acquisition process. Moreover, many images are poor in terms of contrast. Therefore, it is required to suppress the noises and artefacts and enhance the cell contrast. First, to denoise the image, we employed five filtering methods, namely, Gaussian filter, Laplacian filter, Wiener filter, median filter, and mean filter. The peak signal-to-noise ratio (PSNR) for each method is computed. The PSNR is used to assess the quality of the filtered image. The higher the PSNR, the better is the image quality. Figure 4(a) compares the results of the PSNR. We selected the median filter because it exhibits the highest PSNR. The median filter is nonlinear method to suppress the noises by windowing the noisy image. Default window size  $3 \times 3$  is used to remove the small noises. To enhance the cell contrast, three enhancement methods, namely, histogram equalization, intensity adjustment, and CLAHE, are applied. The contrast improvement index (CII) is computed for each method. The CII is utilized for assessing the performance of the image enhancement techniques in terms of the luminance, contrast, and structure. The higher the CII value, the better is the contrast. Figure 4(b) compares the results of the CII. The CLAHE is selected because it results in the highest CII. CLAHE with 8-bit histogram bins is utilized to enhance the contrast of the cell. Figure 5 shows the resulting image after the CLAHE and median filter are applied.

**3.4. Segmentation Stage.** The segmentation step is an important step toward automatic image analysis. This step aims to discriminate between the foreground (the desired object) and the background of the image. The objective of this stage is to extract the cell nuclei from the entire image. In this section, we briefly summarize the twelve segmentation methods. We categorized them into four groups: thresholding, clustering, active contour, and graph-based techniques.

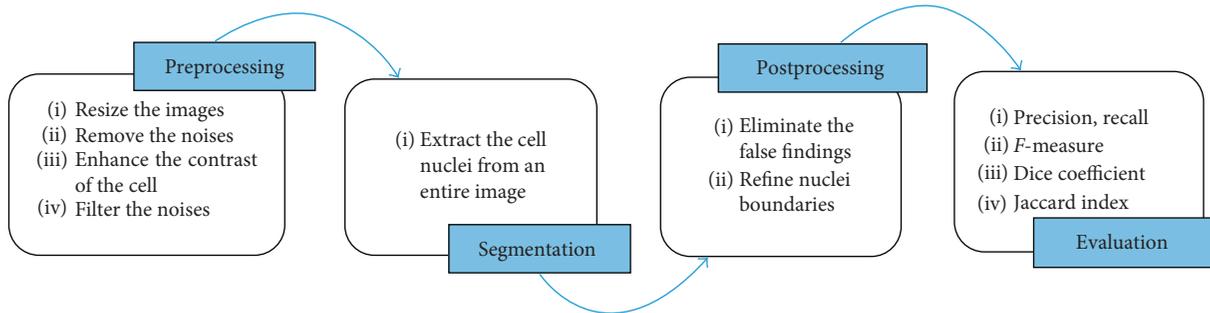


FIGURE 3: Generalized framework of cell nuclei segmentation for cytology pleural effusion images.

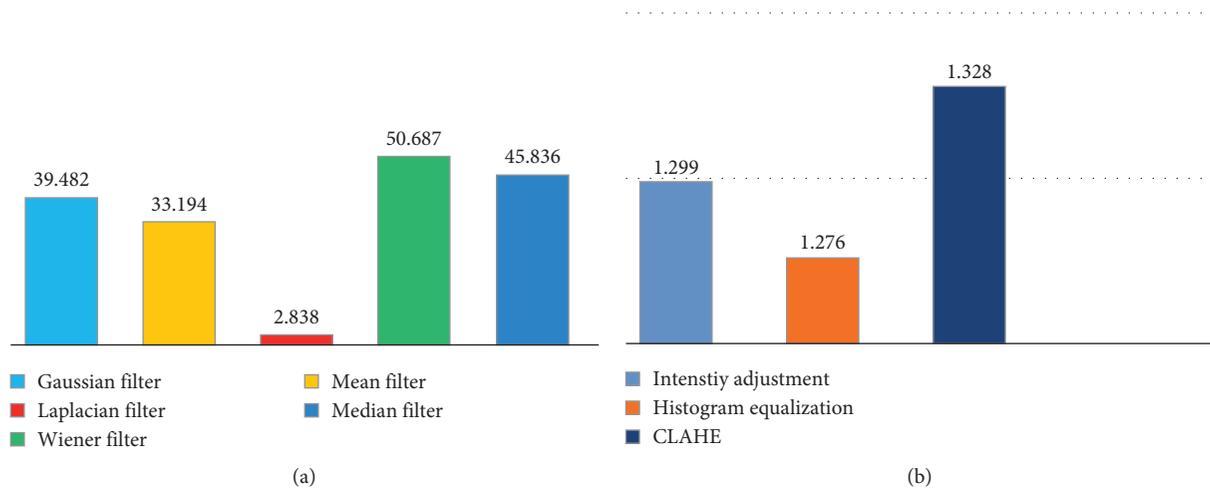


FIGURE 4: Image quality assessment metrics: (a) comparison of filtering methods in terms of peak signal-to-noise ratio (PSNR) and (b) comparison of different contrast enhancement methods in terms of the contrast improvement index (CII).

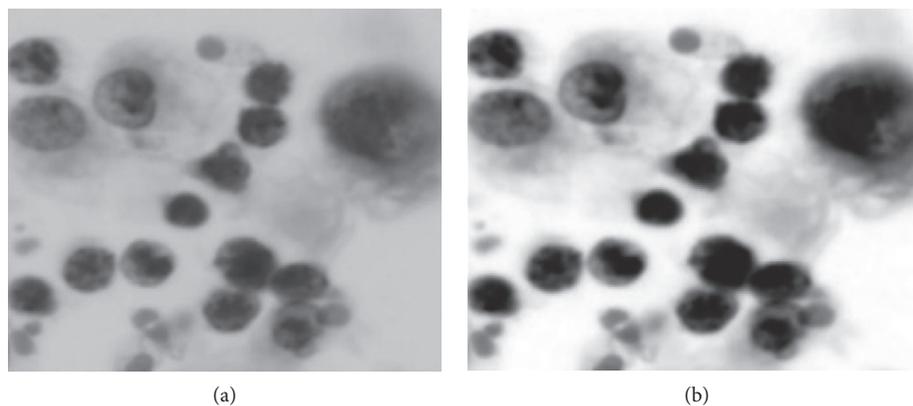


FIGURE 5: Preprocessing stage: (a) grayscale image and (b) preprocessed image after median filter and CLAHE (note that the image was cropped for better visibility).

**3.4.1. Thresholding Techniques.** The thresholding technique is the simplest segmentation method in terms of the gray-level image histogram. It aims to discriminate the foreground and the background by selecting an adequate threshold value. The threshold value can be global or local. A single optimal threshold is utilized for the whole image in the global thresholding, whereas the threshold for each pixel is computed depending on its local properties in the local

approach. Many of thresholding methods take the normalized histogram of the image as the input parameter.

*(1) Otsu's Thresholding Method.* Otsu's method, which is invented by Nobuyuki Otsu, is one of the global thresholding methods. The aim of Otsu's method is to determine the optimal threshold that minimizes the intraclass variance [34]. The algorithm steps are given in Algorithm 1.

- (1) Find the histogram and the probabilities for all intensity levels
- (2) Initiate the class probability ( $w_i$ ) and mean ( $\mu_i$ )
- (3) Move to all possible maximum intensity of thresholds
- (4) Modify  $w_i$  and  $\mu_i$
- (5) Select the maximum value among class variances.

ALGORITHM 1: Otsu's thresholding method.

(2) *Isodata Thresholding Method.* The Isodata thresholding method is one of the global image thresholding methods wherein the following iterative procedure is employed [35]. It requires the initial threshold value as the input. The mean intensity of the image histogram is set as the initial value. The processing steps of Isodata thresholding are given in Algorithm 2.

(3) *Maximum Entropy Thresholding.* The maximum entropy thresholding method is one of the global thresholding methods. Similar to Otsu's method, an optimal threshold is selected in the maximum entropy thresholding method by maximizing the information measured between the object and the background [36]. It takes the normalized histogram of the image as the input parameter. The processing steps are summarized in Algorithm 3.

(4) *Cross Entropy Thresholding.* The cross entropy thresholding is one of the entropy methods. Numerous algorithms have been developed for the cross entropy thresholding. Here, we focus on the one proposed by Li and Lee, which is summarized as follows [37]. Similar to maximum entropy thresholding, cross entropy thresholding takes the histogram of the image as the input parameter. Algorithm 4 describes the processing steps of cross entropy thresholding.

(5) *Fuzzy Entropy Thresholding.* Fuzzy entropy is defined as the measure of uncertainty of a fuzzy set, the procedure of which is given below. It requires two input parameters: (i) histogram of the image to compute the probability distribution and (ii) fuzzy membership function as given in [38]. The algorithm steps are summarized in Algorithm 5.

(6) *Minimum Error Thresholding.* In this method, the image segmentation is based on the average pixel optimization [39]. It requires the normalized histogram of the gray-level image as the input parameter. The idea behind this thresholding technique is presented in Algorithm 6.

(7) *Adaptive Thresholding.* Adaptive thresholding is the most famous local thresholding method for images with uneven illumination. It aims to select threshold values for each region based on its local properties. The local window size is empirically set as 12. We summarized the adaptive thresholding procedures as in Algorithm 7 [40]:

3.4.2. *Clustering Techniques.* Clustering-based segmentation methods aim to group the collection of pixels into clusters.

- (1) Set the initial value of threshold ( $T$ )
- (2) Segment the images into two regions ( $R_1$  and  $R_2$ ) using  $T$  with this formation ( $R_1 < T$  and  $R_2 \geq T$ )
- (3) Find the mean intensity level ( $u_1$ ) for  $R_1$  and ( $u_2$ ) for  $R_2$
- (4) Find new threshold value ( $T = (u_1 + u_2)/2$ )
- (5) Iterate steps 2 to 4 to find the difference in  $T$ . Select  $T$  when it is smaller than the predefined parameter.

ALGORITHM 2: Isodata thresholding method.

- (1) Compute the normalized histogram
- (2) Compute the entropy of white and black pixels
- (3) Select the optimal threshold by maximizing the entropy of white and black pixels.

ALGORITHM 3: Maximum entropy thresholding method.

- (1) Find histogram of the image
- (2) Construct the resulting image by setting the threshold value  $T$
- (3) Compute cross entropy between the original image and the resulting image
- (4) Minimize the cross entropy to optimize the threshold value.

ALGORITHM 4: Cross entropy thresholding method.

- (1) Compute the two probability distributions for foreground and background
- (2) Transform the image into fuzzy set using membership function
- (3) Compute the membership functions for background and foreground using threshold  $T$
- (4) Rewrite the fuzzy form of entropic for step 1
- (5) Obtain the optimal threshold by maximizing the fuzzy entropy.

ALGORITHM 5: Fuzzy entropy thresholding method.

- (1) Compute the histogram that considers the mixture of two normal distributions having respective mean and variance, and respective proportions
- (2) Set the trial threshold value  $T$  for modeling the two resulting pixel populations
- (3) Model the two populations using the normal distribution
- (4) Set different levels as the threshold value
- (5) Compute the fitting criterion for each threshold value
- (6) Select the threshold value which minimizes the fitting criterion as the optimal threshold.

ALGORITHM 6: Minimum error thresholding method.

- (1) Binarize the image with  $T$
- (2) Thin the binary image
- (3) Remove all branch points
- (4) Place the remaining endpoints in line to use as starting point for tracking
- (5) Track the object with  $T$
- (6) Set the criteria  $T = T - 1$  if the object passed testing; otherwise, return to step 5.

ALGORITHM 7: Adaptive thresholding method.

The pixels in the same cluster are closely related to one another.

(1) *K-Means Clustering*. The k-means clustering is one of the clustering methods wherein the data are divided into a specific number of groups by minimizing the within-class variance [41]. The processing steps of k-means clustering based segmentation is presented in Algorithm 8.

(2) *Fuzzy C-Means Clustering*. The fuzzy c-means clustering is one the most popular fuzzy clustering methods, wherein the data are partitioned into two or more fuzzy clusters by maximizing the objective function [42]. Algorithm 9 summarizes the steps involved in this technique.

(3) *Mean Shift Clustering*. Among the clustering-based segmentation methods, the mean shift segmentation is known as an advanced and highly useful technique. In the mean shift, a window is defined for each data point and the mean is subsequently computed. The center of the window is shifted to the mean and the iteration is performed until it converges [43]. Algorithm 10 describes the processing steps of mean shift clustering-based segmentation technique.

3.4.3. *Graph-Based Segmentation Technique*. Graph-based models consider the image as a weighted graph. Every pixel in the image is considered as the node in the graph. The similarities between two nodes are stated as edge weights.

(1) *Min Cut*. A graph cut is a partition of the graph directly or indirectly into two disjoint subsets. The graph is partitioned into clusters using the min cut method. Each cluster is considered as an image segment. The min cut method uses the highly connected subgraph (HCS) algorithm to find the clusters [44]. It can be formulated as follows:

$$\text{cut}(X, Y) = \sum w(i, j), \quad (1)$$

where  $i \in X$ ,  $j \in Y$ , and  $X$  and  $Y$  are two partitioned disjoint sets.

3.4.4. *Active Contour Segmentation Technique*. Active contour models (or snakes) aim to delineate the objects using the energy minimization function. It is performed by assigning the object boundary as the initial contour and subsequently evolving the contour to detect the desired object boundary by driving image forces [45].

- (1) Select  $k$  cluster centers
- (2) For each pixel of an image, find its closest center and assign to the closest class
- (3) Update every center as the mean of its points
- (4) Repeat until it convergence or when there are no changes during the assignment step, or when the average distortion per point decreases slightly
- (5) Reshape the cluster pixels into the image.

ALGORITHM 8: K-means clustering method.

- (1) Choose random centroids, at least two
- (2) Compute the fuzzy membership matrix
- (3) Calculate the cluster center
- (4) Repeat steps 2 and 3 until the minimum objective function value is achieved.

ALGORITHM 9: Fuzzy c-means clustering method.

- (1) Find features (color, gradients, texture, etc.)
- (2) Initialize windows at individual pixel locations
- (3) Perform mean shift for each window until convergence
- (4) Merge windows that end up near the same “peak” or mode.

ALGORITHM 10: Mean shift clustering method.

(1) *Active Contour without Edges (Chan–Vese)*. Among many active contour methods, the active contour without edges, known as the Chan–Vese method, is widely used in cell segmentation. It helps detect the objects without a gradient. It has the ability to segment smoothed contour objects by shrinking the contours and works well on convex objects.

3.5. *Postprocessing Stage*. Postprocessing is an important step to optimize the segmentation results. While most of the segmented regions obtained through the segmentation step will likely correspond to the nuclei regions, there may also be the existence of false findings such as blood cells and artefacts, which must be filtered out. Therefore, it is essential to remove those spurious regions and retain the valid nuclei. A series of the morphological operations is utilized to remedy above problems [46]. Firstly, the morphological filtering method is applied to remove the small objects since the artefacts and blood cells are usually smaller than the actual nuclei. The processing of eliminating the spurious objects is given as Pseudocode 1.

As described in the above pseudocode, it is required to specify the size of  $P$  which is the threshold between the actual nuclei and the spurious regions. The optimal value of  $P$  is empirically set as 1500 pixels. After removing false findings, we further applied the morphological closing and opening operations for nuclei shape’s refinement and simplification. The structuring element (SE) with disk shape and radius ( $R$ ) is used.  $R$  is set as 5 and 12 for opening and

```

Input: Candidate nuclei regions (Candidatenuclei)
Output: Actual nuclei regions (Actualnuclei)
(1) Determine the connected components using 8-connectivity
(2) Count the number of components ( $N$ )
(3) Compute the area of each component  $A_n, n \in N$ 
(4) Remove small objects using the predetermined value ( $P$ ) as follows:

  For  $i = 1 : N$ 
    If  $A_i \geq P$ 
      NucleiMask =  $A_i$ ;
    End
  End

```

PSEUDOCODE 1: Pseudocode for removing the false findings.

closing, respectively. Equations (2) and (3) formulate opening and closing operations, respectively:

$$\text{actual}_{\text{nuclei}} \cdot \text{SE} = (\text{actual}_{\text{nuclei}} \ominus \text{SE}) \oplus \text{SE}, \quad (2)$$

$$\text{opened}_{\text{nuclei}} \cdot \text{SE} = (\text{opened}_{\text{nuclei}} \oplus \text{SE}) \ominus \text{SE}, \quad (3)$$

where  $\oplus$  and  $\ominus$  represent the dilation and erosion, respectively. The sampled visual results of before and after preprocessing are depicted in Figure 6.

## 4. Benchmark Setting

**4.1. Experimental Results.** This study was carried out using MATLAB (2013 version) on a computer with an Intel Core i7 processor clocked at 2.50 GHz and with 8 GB of RAM. A local dataset of 35 cytology images of pleural effusion and its ground truth images are used. In this study, we considered three main stages to extract the cell nuclei. The first stage is used to deal with the image quality. We employed different enhancement and filtering methods. The image quality assessment metrics, namely, CII and PSNR, are computed to select the best ones. Based on the CII and PSNR results, we

selected CLAHE and median filtering methods to enhance the image quality. The segmentation stage is the most important stage in extracting the cell nuclei regions. We experimentally employed twelve segmentation methods, as explained in Section 3. Finally, the preprocessing stage is performed to refine the boundaries of the segmented cell nuclei and remove the undesired regions using morphological operations. As the segmentation stage is paramount, the segmentation results vary depending on the segmentation methods. Figure 7 shows the visual results of the cell nuclei segmented using different segmentation methods.

To quantitatively evaluate the segmentation methods, five pixel-based performance metrics, namely, precision, recall, measure, Jaccard Index (JI), and Dice similarity coefficient (DSC), were computed for each algorithm. The examined methods are evaluated by comparing with hand-drawn ground truth images. Each connected region in the segmented results is considered as one nucleus while ignoring the number of nucleus inside the region. Ground truth images are also prepared the same way. Figure 7(b) depicts the sample of ground truth image. The performance measures can be formulated as follows:

$$\begin{aligned} \text{precision (Pre)} &= \frac{\text{true positive}}{\text{true positive} + \text{false positive}}, \\ \text{recall (Re)} &= \frac{\text{true positive}}{\text{true positive} + \text{false negative}}, \\ \text{Fmeasure (Fm)} &= \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}, \\ \text{JI} &= \frac{\text{true positive}}{\text{true positive} + \text{false positive} + \text{false negative}}, \\ \text{DSC} &= \frac{2 * \text{true positive}}{2 * \text{true positive} + \text{false positive} + \text{false negative}}. \end{aligned} \quad (4)$$

The performance metrics were used to quantitatively evaluate the cell nuclei in the segmented image individually. Table 1 lists the evaluation results. The compared quantitative

results show that the segmentation performances of Otsu's method, k-means, mean shift, Chan-Vese level set method, and graph-based min cut are excellent. The accuracies meet

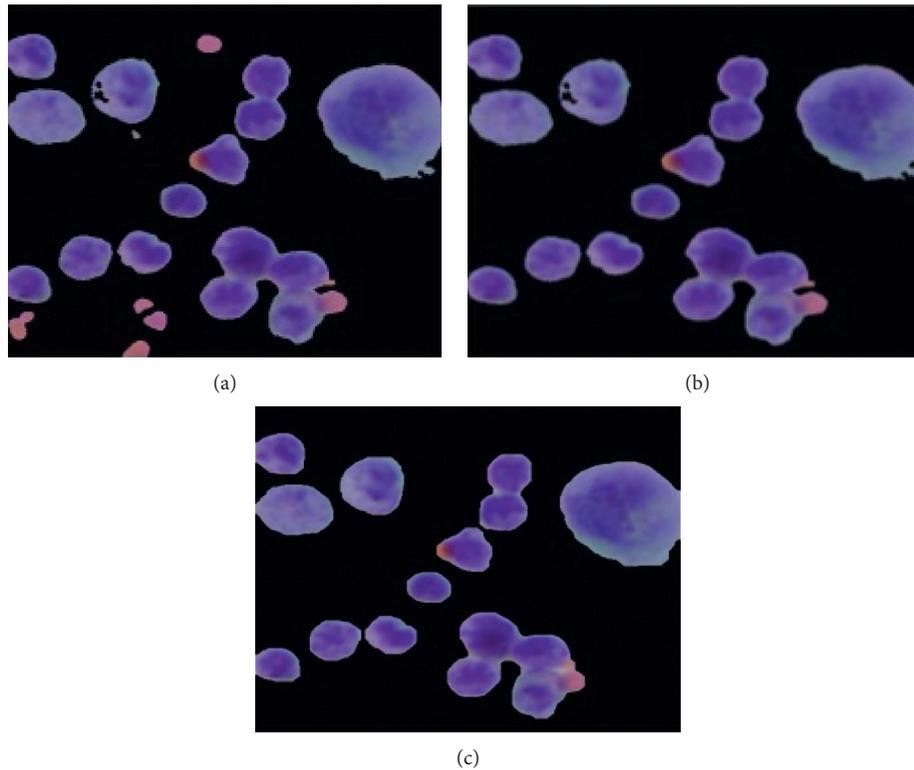


FIGURE 6: The sample visual results of before and after postprocessing stage: (a) candidate nuclei extracted using Otsu thresholding, (b) after removing spurious objects using morphological filtering, and (c) after refining the contours of the nuclei using morphological opening and closing.

clinical requirements. For the highlighted methods, we further evaluated the nuclei detection rate (NDR) of the images depending on the recall value. The NDR is considered as true positive when the recall is greater than 60%. The overall NDR of each algorithm is computed and compared, as shown in Figure 8(a). Similar to the NDR, we estimated the abnormal NDR. Figure 8(b) shows the comparison results. The comparison results show that the mean shift clustering method exhibits the best performance in terms of the overall NDR and abnormal NDR. To evaluate the time complexity of each method, the computational time of each method is computed and compared, as shown in Figure 9. Otsu's method is found to be relatively simple and fast. In contrast, the Chan-Vese method is computationally expensive.

**4.2. Parameters Tuning and Discussion.** The highlighted segmentation methods are discussed along with their adjustable parameters, advantages, and limitations. The experiment results show that the performances of the segmentation methods strongly depend on the tuning parameters. Therefore, it is required to properly select the most relevant parameters for our applications. We experimentally set and adjusted different parameters in each segmentation method and selected the most effective one for all images.

As Otsu's method is a nonparametric method, it is not required to specially assign prior parameters. However, Otsu's method is sensitive to outliers. To deal with this issue, the CLAHE and median filter methods are employed to enhance

the image quality before applying Otsu's method. The cytology pleural effusion images comprise three main parts: cell nuclei, cytoplasm and blood cells, and background. As the color of the nuclei region appears to be dark purple, with other parts appearing lighter in color, the image histogram is assumed to be a bimodal distribution. Thus, Otsu's method provides relatively good performance in our application. Otsu's method is relatively simple and the result is promising. Therefore, it can be applied to real applications. However, the performance is degraded when the image contains significant noises because the method is sensitive to noise.

The segmentation result of the  $k$ -means clustering method strongly depends on initializing the  $k$  clusters. A poor initialization can significantly affect the clustering performance and result in a poor convergence speed. Therefore, we set multiple  $k$  clusters for the test and chose the most effective one. The numbers of  $k$  clusters were set as 2, 3, 4, and 5. When  $k$  is 3, the nuclei are mixed with other image components. Thus, it was difficult to separate only the nuclei regions from the mixed clustered regions. When  $k$  is 4, 5, or more, the nuclei are broken into multisegmented images because of the high variation in the pixel intensity within the nuclei. Hence, the nuclei regions should be obtained from multiple images. When  $k$  is 2, the nuclei are segmented in a straightforward manner. In addition, the nuclei appear dark in color with some regions remaining bright. This fact also supports in setting up the value of  $k$  as 2 to cluster two groups (dark and bright colors). Figure 10 shows the visual segmented cell nuclei with different  $k$

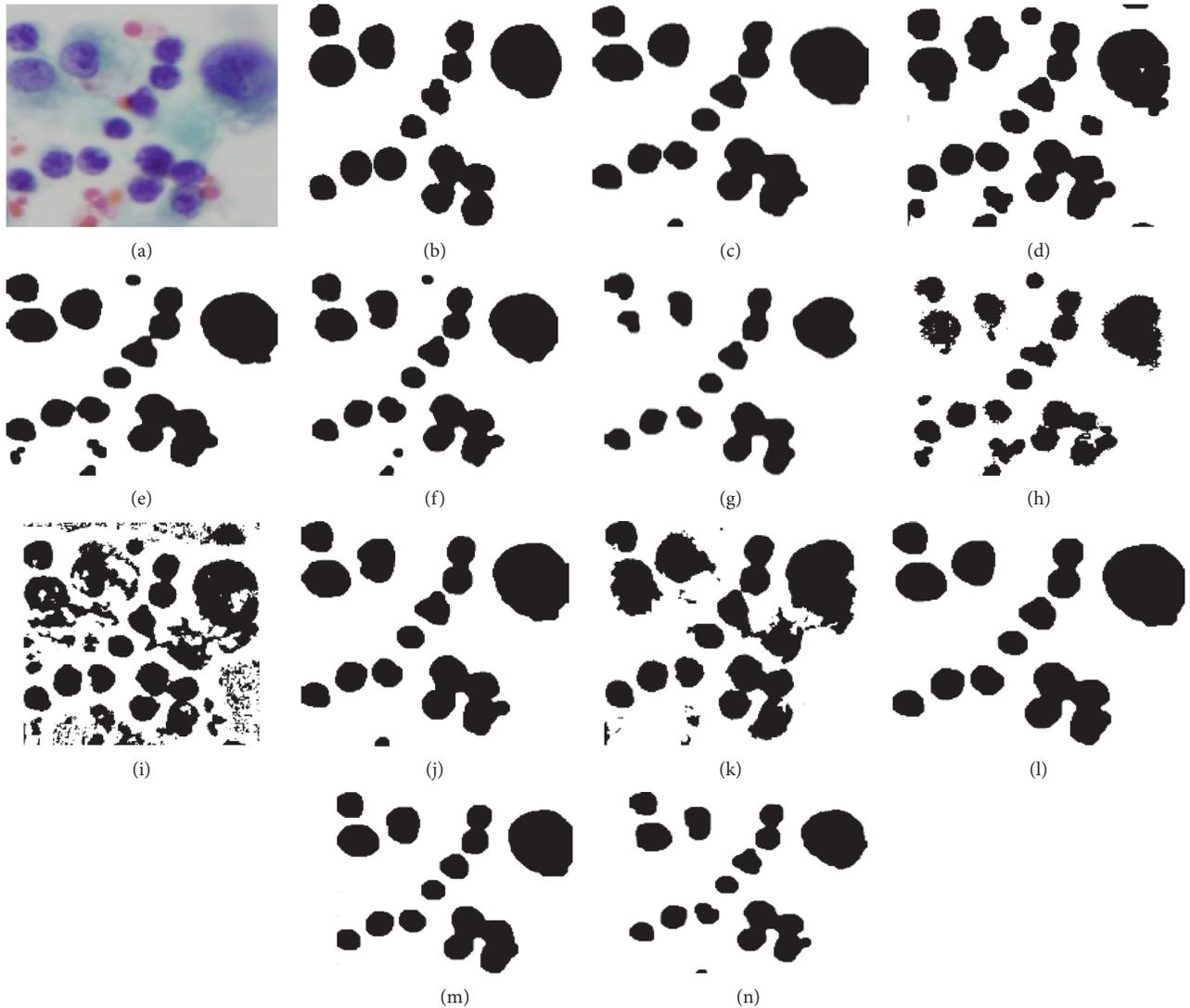


FIGURE 7: Visual cell nuclei segmented using twelve segmentation methods: (a) original image, (b) ground truth image, (c) Otsu's method, (d) Isodata, (e) maximum entropy, (f) cross entropy, (g) minimum error, (h) fuzzy entropy, (i) adaptive thresholding, (j) k-means clustering, (k) fuzzy c-means clustering, (l) mean shift clustering, (m) Chan-Vese method, and (n) graph-based min cut (note that the images were cropped for better visibility here).

TABLE 1: Quantitative experimental results.

Methods/evaluation	Pre	Re	Fm	JI	DSC (%)
Otsu's method	0.91	0.89	0.90	0.89	94
Isodata	0.86	0.84	0.85	0.84	91
Maximum entropy	0.84	0.94	0.88	0.94	91
Cross entropy	0.94	0.82	0.87	0.82	90
Minimum error	0.85	0.82	0.83	0.82	89
Fuzzy entropy	0.68	0.67	0.68	0.67	80
Adaptive thresholding	0.96	0.50	0.66	0.50	67
k-means	0.90	0.89	0.89	0.89	94
Fuzzy c-means	0.94	0.77	0.85	0.77	77
Mean shift	0.93	0.91	0.92	0.91	95
Chan-Vese method	0.89	0.87	0.88	0.87	94
Graph-based min cut	0.87	0.95	0.91	0.87	93

clusters. Moreover, it is worth noticing that the k-means clustering method performs well for round-shaped objects. Hence, the method is effective in segmenting cell nuclei

because the cell nuclei are largely round shaped. In addition, it is simple, fast, and easy to implement. However, the disadvantage is that the k-means clustering is extremely sensitive to the  $k$  clusters and performs badly when the clusters are convex shaped.

In contrast to the k-means clustering method, the mean shift is a nonparametric clustering method. It is not necessary to define the clusters and restrict the cluster shape. Nevertheless, the clustering result of the mean shift strongly depends on the bandwidth size. It is required to carefully select the most relevant size for particular applications. In our experiments, we experimentally set multiple bandwidth sizes and chose the best one. Figure 11 shows the significant differences in the segmentations in terms of the bandwidth sizes. The experiment results show that a bandwidth size of 0.2 exhibits the best clustering performance, appropriate for cell nuclei segmentation. The advantage of the mean shift is that it is not necessary to initialize

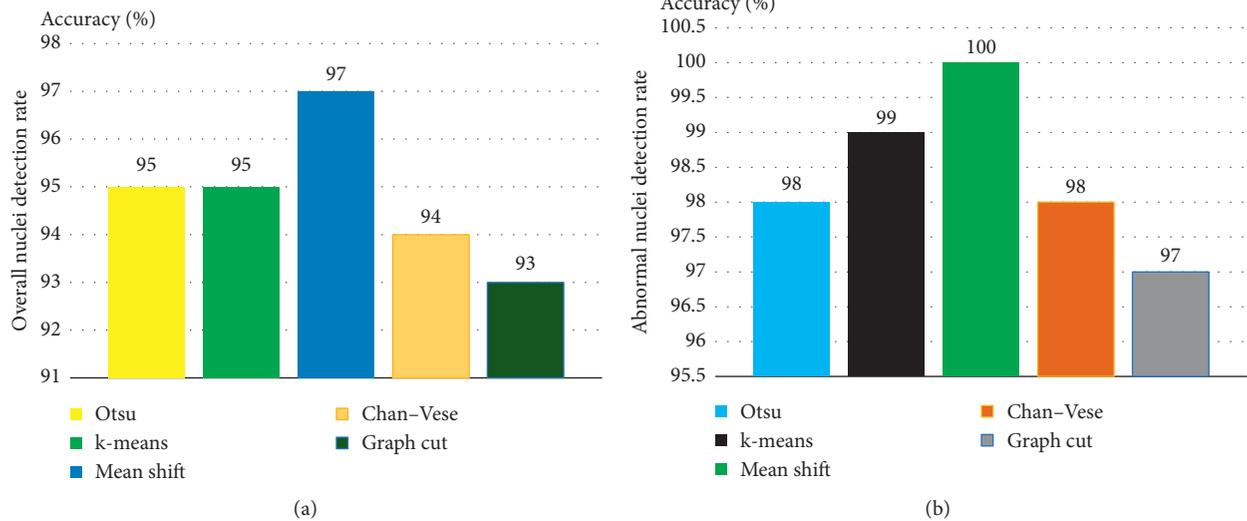


FIGURE 8: Comparison of nuclei detection rates in terms of the recall value: (a) overall nuclei detection rate and (b) abnormal cell nuclei detection rate.

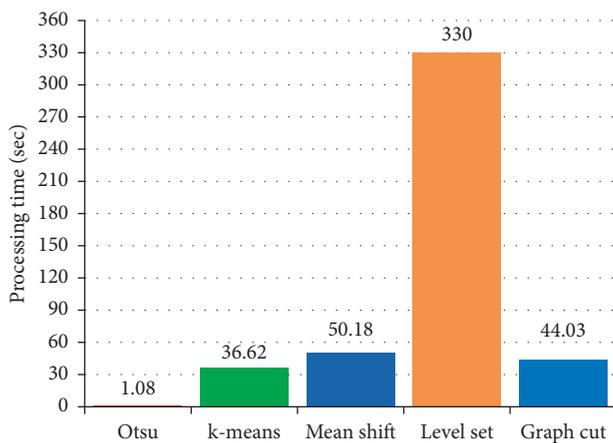


FIGURE 9: Processing time of five highlighted methods.

the cluster numbers; moreover, the method need not be robust to outliers. Only the size of the bandwidth is required to be set. The limitation of the mean shift is that it is computationally expensive compared to other clustering methods, as many windows need to be shifted, thus making many computations redundant.

In the Chan-Vese level set method, the boundaries of the regions are used as a mask, which is initial contour evolution of the segmentation start. To achieve a fast and accurate output, we initially specified the mask that is close to the nuclei regions. The mask either shrinks or expands based on the image features. In addition, it is crucial to specify an appropriate maximum iteration for the contour evolution. The iteration is stopped if the maximum iteration is reached, when the energy remains constant, or when the contour is not moving. The maximum number of iterations affects the largest variation in the segmentation results. Therefore, we experimentally tested with different iterations and chose the most effective one for all images. However, it is the main

limitation, as the images contain various types of cells. It is difficult to fix the number of iterations if more images are added into the datasets. If a large number of maximum iterations are set, the computation becomes expensive. In contrast, a small number of iterations lead to undersegmentation, because the iteration is stopped before finishing the contour evolution. Figure 12 depicts the segmentation results obtained through different iteration numbers.

In the graph-based min cut segmentation, two parameters need to be defined. The first one is alpha, which is the penalty parameter with respect to the total variation term. For the case wherein the image-edge weights are incorporated, alpha is given by the constant in all cases. For the case with image-edge weights, alpha is given using the two pixel-wise weighted functions. The second parameter is the step size of the augmented Lagrangian method, the optimal range of which is (0.3, 3). We set it as 0.3, as it is not significantly different for segmentation. A significant variation in the segmentation result is found when setting up different values of alpha. We experimentally tested with different alpha values and chose the best one. Figure 13 depicts the visual result of the segmentation in terms of the alpha values. We chose 0.3 as the alpha value in our application. The graph-based min cut method is simple, easy to control, and fast in processing. It returns the clusters as image segments. However, the drawback is that multiple small segments may be separated by cutting small sets of isolated nodes in the graph.

Apart from the highlighted methods, the accuracies of the maximum entropy, Isodata, cross entropy, and minimum error thresholding methods are very close to real requirements. There is a high possibility that the accuracy can be further improved by adding more effective pre- and postprocessing steps. However, the fuzzy entropy, adaptive thresholding, and fuzzy c-means clustering methods exhibit very low accuracy. The accuracies of these methods can be improved by combining them with appropriate

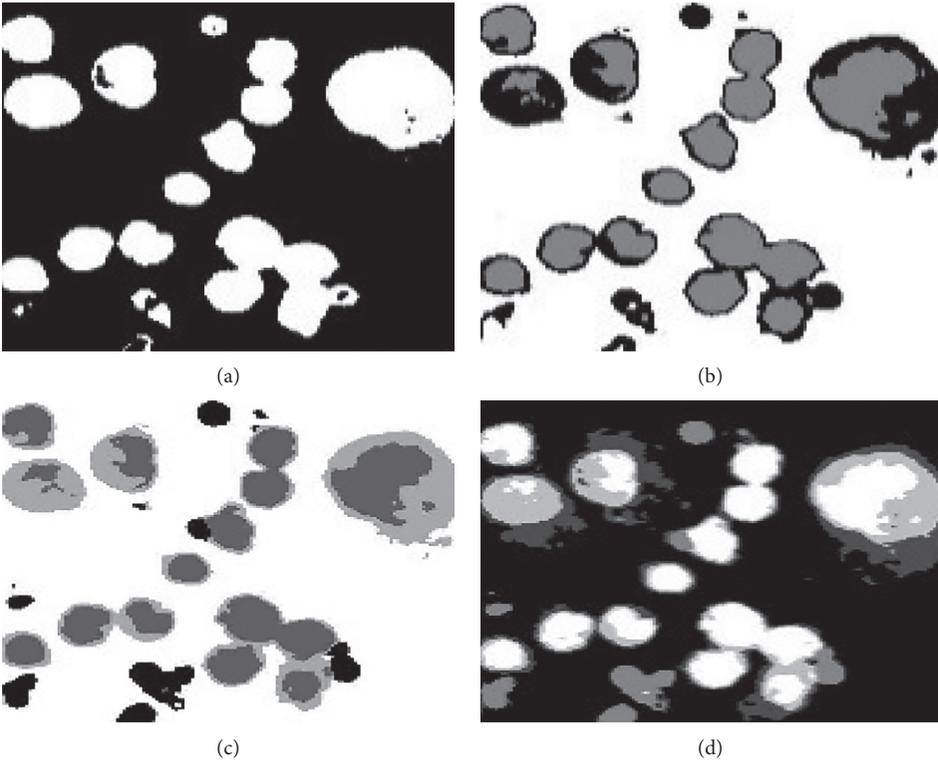


FIGURE 10: Image index labeled with different  $k$  clusters: (a)  $k = 2$ , (b)  $k = 3$ , (c)  $k = 4$ , and (d)  $k = 5$ .

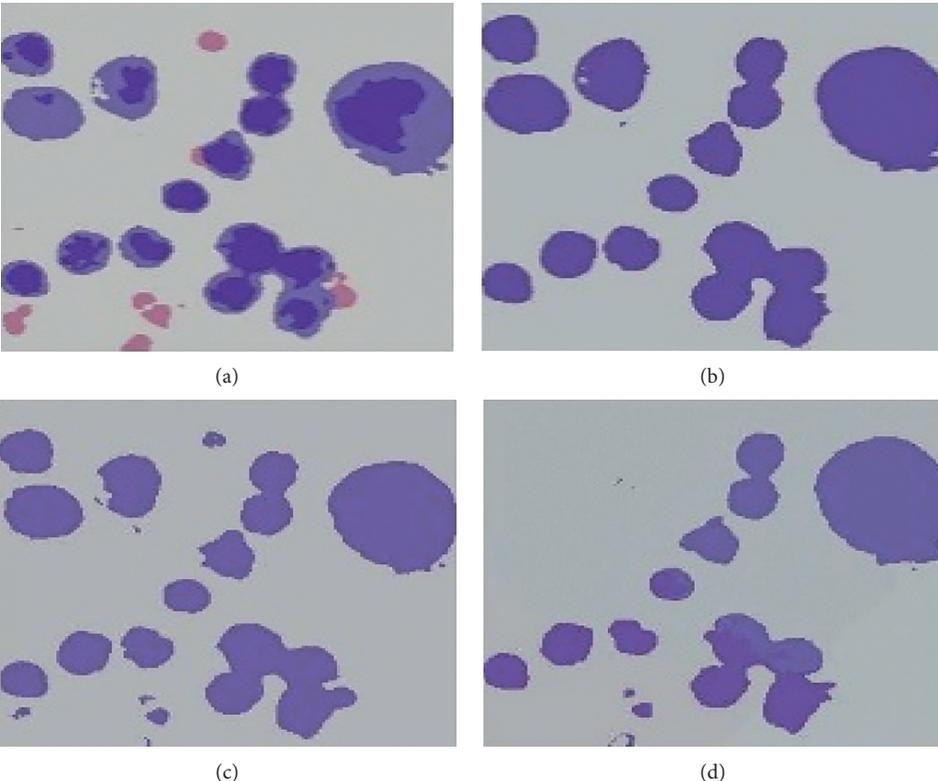


FIGURE 11: Different clustering results in terms of different bandwidth (bw) sizes: (a)  $bw = 0.1$ , (b)  $bw = 0.2$ , (c)  $bw = 0.3$ , and (d)  $bw = 0.4$ .

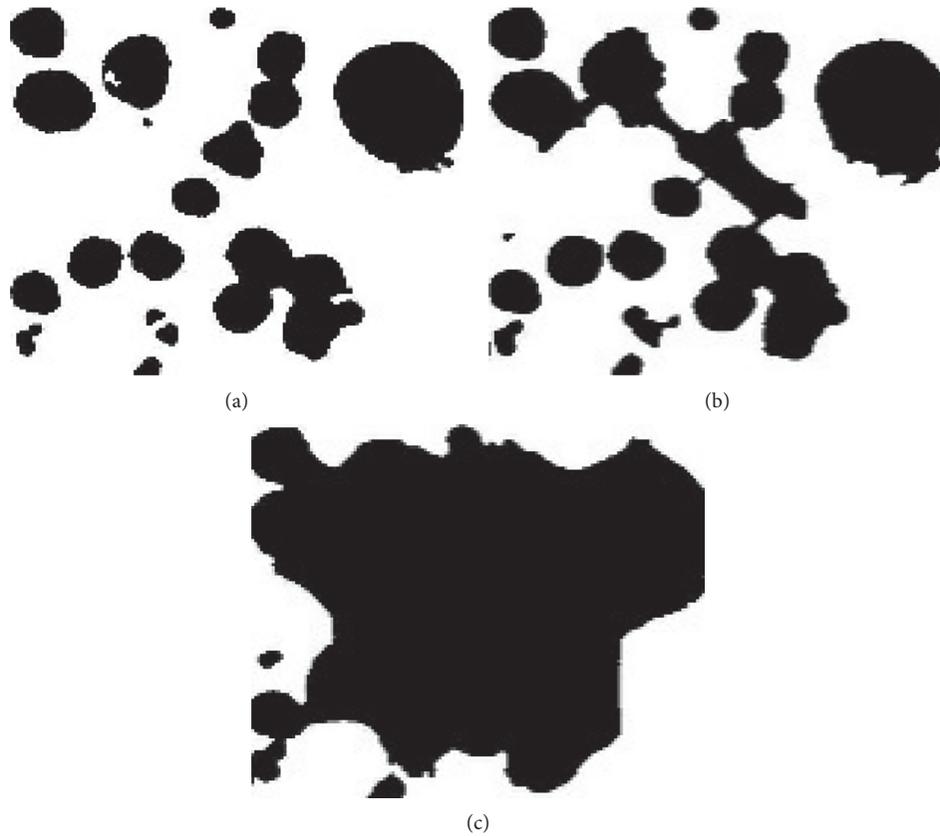


FIGURE 12: Segmentation results obtained using different iterations: (a) iterations = 500, (b) iterations = 300, and (c) iterations = 100 (note that iterations were tested on cropped region of image for better comparison, not on the real input images).

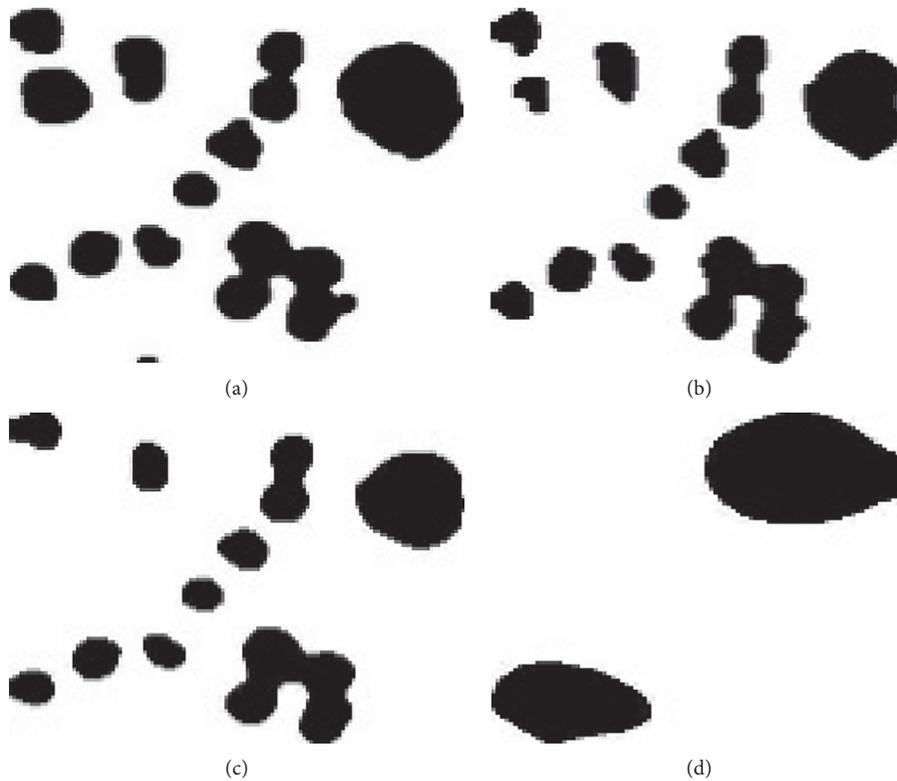


FIGURE 13: Variation in segmentation results in terms of different alpha values (av): (a) av = 0.3, (b) av = 0.5, (c) av = 10, and (d) av = 15.

segmentation methods. In the future, integrated or hybrid segmentation methods will be studied. This study covers only the extraction of the nuclei regions from surrounding objects in the entire image. The examined methods are not capable of splitting the overlapping cells. It should be noted that the touching or overlapping or clustering cells can be isolated into individual ones in further stages. In order to do so, these methods may integrate with the overlapped object splitting techniques such as the watershed methods, contour concavity methods, rule-based splitting methods, bottleneck method, and so on.

## 5. Conclusions

In this paper, we presented twelve individual image segmentation algorithms to extract the cell nuclei from cytology pleural effusion images. Each method includes three stages: preprocessing, segmentation, and postprocessing. The accuracies (DSC) of the five segmentation methods, namely, Otsu's method, k-means clustering, mean shift, Chan–Vese method, and graph cut method, were 94, 94, 95, 94, and 93%, respectively; the average computational times for an image were 1.08, 36.62, 50.18, 330, and 44.03 seconds, respectively. The results meet clinical requirements. Therefore, they can be practically used as a prerequisite step in developing CAD systems. The Chan–Vese method is computationally expensive compared to others. In contrast, Otsu's method is relatively simple and fast. It is our hope that the results and observations will be useful for current and potential future studies on cytology images of pleural effusion. Unfortunately, the examined methods are not capable of separating the overlapping nuclei. In further study, they need to integrate with additional splitting methods to isolate the overlapping nuclei into individual ones. As part of the future work, more state-of-the-art segmentation methods will be explored. In addition, it would be interesting to combine different segmentation methods for improving the segmentation accuracy; however, this requires research efforts beyond the scope of this paper. It would be worthwhile to study the implementation and combination of different algorithms. Our ultimate goal is to implement an effective CAD system for malignant pleural effusion.

## Conflicts of Interest

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

## Acknowledgments

The authors would like to express gratitude to AUN/SEED-Net for their financial support. The authors also greatly acknowledge the Department of Pathology, Faculty of Medicine, Srinakharinwirot University, Thailand, for the insightful suggestions, including their cooperation for the dataset and ground truth segmentation.

## References

- [1] World Health Organization, *Cancer*, WHO, Geneva, Switzerland, <http://www.who.int/news-room/fact-sheets/detail/cancer>.

- [2] J. Ferlay, I. Soerjomataram, M. Ervik et al., *Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11, GLOBOCAN v1.0*, IARC, Lyon, France, 2012.
- [3] Pleural effusion,” <http://www.macmillan.org.uk>.
- [4] M. E. Roberts, E. Neville, R. G. Berrisford et al., “Management of a malignant pleural effusion: British Thoracic Society pleural disease guideline 2010,” *Thorax*, vol. 65, no. S2, pp. ii32–ii40, 2010.
- [5] C. Jung and C. Kim, “Segmenting clustered nuclei using H-minima transform-based marker extraction and contour parameterization,” *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 10, pp. 2600–2604, 2010.
- [6] J. P. Vink, M. B. Van Leeuwen, C. H. M. Van Deurzen, and G. De Haan, “Efficient nucleus detector in histopathology images,” *Journal of Microscopy*, vol. 249, no. 2, pp. 124–135, 2013.
- [7] M. Veta, P. J. van Diest, R. Kornegoor, A. Huisman, M. A. Viergever, and J. P. Pluim, “Automatic nuclei segmentation in H&E stained breast cancer histopathology images,” *PLoS One*, vol. 8, no. 7, Article ID e70221, 2013.
- [8] M. Paramanandam, M. O’Byrne, B. Ghosh et al., “Automated segmentation of nuclei in breast cancer histopathology images,” *PLoS One*, vol. 11, no. 9, Article ID e0162053, 2016.
- [9] Y. Al-Kofahi, W. Lassoued, W. Lee, and B. Roysam, “Improved automatic detection and segmentation of cell nuclei in histopathology images,” *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 4, pp. 841–852, 2010.
- [10] Y. M. George, B. M. Bagoury, H. H. Zayed, and M. I. Roushdy, “Automated cell nuclei segmentation for breast fine needle aspiration cytology,” *Signal Processing*, vol. 93, no. 10, pp. 2804–2816, 2013.
- [11] L. Zhang, H. Kong, C. T. Chin et al., “Segmentation of cytoplasm and nuclei of abnormal cells in cervical cytology using global and local graph cuts,” *Computerized Medical Imaging and Graphics*, vol. 38, no. 5, pp. 369–380, 2014.
- [12] Y. Song, L. Zhang, S. Chen, D. Ni, B. Lei, and T. Wang, “Accurate segmentation of cervical cytoplasm and nuclei based on multiscale convolutional network and graph partitioning,” *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 10, pp. 2421–2433, 2015.
- [13] J. Zhang, Z. Hu, G. Han, and X. He, “Segmentation of overlapping cells in cervical smears based on spatial relationship and overlapping transcurrency light transmission model,” *Pattern Recognition*, vol. 60, pp. 286–295, 2016.
- [14] M. E. Plissiti, C. Nikou, and A. Charchanti, “Automated detection of cell nuclei in pap smear images using morphological reconstruction and clustering,” *IEEE Transactions on Information Technology in Biomedicine*, vol. 15, no. 2, pp. 233–241, 2011.
- [15] Y. Song, E. L. Tan, X. Jiang et al., “Accurate cervical cell segmentation from overlapping clumps in pap smear images,” *IEEE Transactions on Medical Imaging*, vol. 36, no. 1, pp. 288–300, 2017.
- [16] B. Nielsen, F. Albrechtsen, and H. E. Danielsen, “Automatic segmentation of cell nuclei in Feulgen-stained histological sections of prostate cancer and quantitative evaluation of segmentation results,” *Cytometry Part A*, vol. 81, no. 7, pp. 588–601, 2012.
- [17] D. L. L. De Oliveira, M. Z. Do Nascimento, L. A. Neves, M. F. De Godoy, P. F. F. De Arruda, and D. de Santi Neto, “Unsupervised segmentation method for cuboidal cell nuclei in histological prostate images based on minimum cross entropy,” *Expert Systems with Applications*, vol. 40, no. 18, pp. 7331–7340, 2013.

- [18] Z. Liu, J. Liu, X. Xiao et al., "Segmentation of white blood cells through nucleus mark watershed operations and mean shift clustering," *Sensors*, vol. 15, no. 9, pp. 22561–22586, 2015.
- [19] O. Sarrafzadeh and A. M. Dehnavi, "Nucleus and cytoplasm segmentation in microscopic images using K-means clustering and region growing," *Advanced Biomedical Research*, vol. 4, no. 174, 2015.
- [20] C. Reta, L. Altamirano, J. A. Gonzalez et al., "Segmentation and classification of bone marrow cells images using contextual information for medical diagnosis of acute leukemia," *PLoS One*, vol. 10, no. 6, Article ID e0130805, 2015.
- [21] F. Taher, N. Werghi, H. Al-Ahmad, and C. Donner, "Extraction and segmentation of sputum cells for lung cancer early diagnosis," *Algorithms*, vol. 6, no. 3, pp. 512–531, 2013.
- [22] C. Liu, F. Shang, J. A. Ozolek, and G. K. Rohde, "Detecting and segmenting cell nuclei in two-dimensional microscopy images," *Journal of Pathology Informatics*, vol. 7, no. 1, 2016.
- [23] V. R. Korde, H. Bartels, J. Ranger-Moore, and J. Barton, "Automatic segmentation of cell nuclei in bladder and skin tissue for karyometric analysis," in *Proceedings of the European Conference on Biomedical Optics*, Munich, Germany, June 2007.
- [24] J. Kong, L. Cooper, T. Kurc, D. Brat, and J. Saltz, "Towards building computerized image analysis framework for nucleus discrimination in microscopy images of diffuse glioma," in *Proceedings of Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 6605–6608, Boston, MA, USA, August 2011.
- [25] L. Zhang, Q. Wang, and J. Qi, "Research based on fuzzy algorithm of cancer cells in pleural fluid microscopic images recognition," in *Proceedings of International Conference on Intelligent Information Hiding and Multimedia*, pp. 211–214, Kaohsiung, Taiwan, November 2006.
- [26] F. Chen, J. Xie, H. Zhang, and D. Xia, "A technique based on wavelet and morphology transform to recognize the cancer cell in pleural effusion," in *Proceedings of International Workshop on Medical Imaging and Augmented Reality*, pp. 199–203, Shatin, Hong Kong, June 2001.
- [27] K. Y. Win and S. Choomchuay, "Automated segmentation of cell nuclei in cytology pleural fluid images using OTSU thresholding," in *Proceedings of International Conference on Digital Arts, Media and Technology (ICDAMT)*, pp. 14–18, Chiang Mai, Thailand, March 2017.
- [28] K. Y. Win, S. Choomchuay, and K. Hamamoto, "Automated segmentation and isolation of touching cell nuclei in cytopathology smear images of pleural effusion using distance transform watershed method," in *Proceedings of Second International Workshop on Pattern Recognition*, vol. 10443, Singapore, May 2017.
- [29] K. Wu, D. Gauthier, and M. D. Levine, "Live cell image segmentation," *IEEE Transactions on Biomedical Engineering*, vol. 42, no. 1, pp. 1–12, 1995.
- [30] H. Refai, L. Li, and T. K. Teague, "Automatic count of hepatocytes in microscopic images," in *Proceedings of International Conference on Image Processing*, vol. 2, pp. 1101–1104, Barcelona, Spain, September 2003.
- [31] N. Salem, N. M. Sobhy, and M. El Dosoky, "A comparative study of white blood cells segmentation using Otsu threshold and watershed transformation," *Journal of Biomedical Engineering and Medical Imaging*, vol. 3, no. 3, p. 15, 2016.
- [32] A. Mouelhi, M. Sayadi, F. Fnaiech, K. Mrad, and K. B. Romdhane, "Automatic image segmentation of nuclear stained breast tissue sections using color active contour model and an improved watershed method," *Biomedical Signal Processing and Control*, vol. 8, no. 5, pp. 421–436, 2013.
- [33] P. Bamford and B. Lovell, "Unsupervised cell nucleus segmentation with active contours," *Signal Processing*, vol. 71, no. 2, pp. 203–213, 1998.
- [34] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.
- [35] F. R. Velasco, *Thresholding Using the ISODATA Clustering Algorithm (No. CSC-TR-751)*, Maryland Univ College Park Computer Science Center, College Park, MD, USA, 1979.
- [36] J. N. Kapur, P. K. Sahoo, and A. K. Wong, "A new method for gray-level picture thresholding using the entropy of the histogram," *Computer Vision, Graphics, and Image Processing*, vol. 29, no. 3, pp. 273–285, 1985.
- [37] C. H. Li and C. K. Lee, "Minimum cross entropy thresholding," *Pattern Recognition*, vol. 26, no. 4, pp. 617–625, 1993.
- [38] O. J. Tobias and R. Seara, "Image segmentation by histogram thresholding using fuzzy sets," *IEEE Transactions on Image Processing*, vol. 11, no. 12, pp. 1457–1465, 2002.
- [39] J. Kittler and J. Illingworth, "Minimum error thresholding," *Pattern Recognition*, vol. 19, no. 1, pp. 41–47, 1986.
- [40] D. Bradley and G. Roth, "Adaptive thresholding using the integral image," *Journal of Graphics Tools*, vol. 12, no. 2, pp. 13–21, 2007.
- [41] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: analysis and implementation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 881–892, 2002.
- [42] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering algorithm," *Computers and Geosciences*, vol. 10, no. 2–3, pp. 191–203, 1984.
- [43] Y. Cheng, "Mean shift, mode seeking, and clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 8, pp. 790–799, 1995.
- [44] T. F. Chan and L. A. Vese, "Active contours without edges," *IEEE Transactions on Image Processing*, vol. 10, no. 2, pp. 266–277, 2001.
- [45] S. Vicente, V. Kolmogorov, and C. Rother, "Graph cut based image segmentation with connectivity priors," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, Anchorage, AK, USA, June 2008.
- [46] R. M. Haralick, S. R. Sternberg, and X. Zhuang, "Image analysis using mathematical morphology," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 4, no. 4, pp. 532–550, 1987.

## Research Article

# The Cell Research Trends of Asthma: A Stem Frequency Analysis of the Literature

Wenchao Tang <sup>1</sup>, Yi Shang <sup>2</sup>, Bin Xiao <sup>1</sup>, Peitong Wen <sup>1</sup>, Ruoyun Lyu <sup>1</sup>,  
and Ke Ning <sup>3</sup>

<sup>1</sup>School of Acupuncture-Moxibustion and Tuina, Shanghai University of Traditional Chinese Medicine, Shanghai 201203, China

<sup>2</sup>Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO 65211, USA

<sup>3</sup>School of Basic Medical Sciences, Shanghai University of Traditional Chinese Medicine, Shanghai 201203, China

Correspondence should be addressed to Wenchao Tang; [vincent.tang@shutcm.edu.cn](mailto:vincent.tang@shutcm.edu.cn)

Received 15 November 2017; Revised 26 March 2018; Accepted 22 July 2018; Published 23 August 2018

Academic Editor: Weidong Cai

Copyright © 2018 Wenchao 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.

**Objective.** This study summarized asthma literature indexed in the Medical Literature Analysis and Retrieval System Online (MEDLINE) and explored the history and present trends of asthma cell research by stem frequency ranking to forecast the prospect of future work. **Methods.** Literature was obtained from MEDLINE for the past 30 years and divided into three groups by decade as the retrieval time. The frequency of stemmed words in each group was calculated using Python with Apache Spark and the Natural Language Tool Kit for ranking. The unique stems or shared stems of 3 decades were summarized. **Results.** A total of 1331, 4393, and 7215 records were retrieved from 3 decades chronologically, and the stem ranking of the top 50 were listed by frequency. The number of stems shared with 3 decades was 26 and with the first and last 2 decades was 5 and 13. **Conclusions.** The number of cell research studies of asthma has increased rapidly, and scholars have paid more attentions on experimental research, especially on mechanistic research. Eosinophils, mast cells, and T cells are the hot spots of immunocyte research, while epithelia and smooth muscle cells are the hot spots of structural cell research. The research trend is closely linked with the development of experimental technology, including animal models. Early studies featured basic research, but immunity research has dominated in recent decades. The distinct definition of asthma phenotypes associated with genetic characteristics, immunity research, and the introduction of new cells will be the hot spots in future work.

## 1. Introduction

Asthma is a major public health problem around the world, affecting individuals across the age spectrum from infants to older adults. Therefore, research on its pathogenesis and treatment has been a hot topic in the study of respiratory diseases. It is now well accepted that cell activity has a close relationship with pathogenesis of asthma, and numerous basic and clinical studies focus on different types of relevant cells.

As a typical example of type I hypersensitivity, the research of immune cells concerned with asthma, such as lymphocytes, monocytes, and mast cells, is most common. For instance, T cell subsets include CD4<sup>+</sup>, CD8<sup>+</sup> T cells [1], Th2, Th17 [2], Th9 [3], and so on. Except for immune-relevant cells, structural cells

of the airways, such as epithelial cells [4], smooth muscle cells [5], and bronchial myofibroblasts [6], are also an important focus of research. In recent years, cell therapy has attracted the attention of researchers to treat asthma and its complications. A study revealed that bone marrow-derived mesenchymal stem cell (BMSC) therapy significantly suppressed lung pathology and inflammation in the ovalbumin-induced asthma mouse model [7].

Currently, there has been continued interest in targeting airway cells for developing new asthma treatments. Therefore, it has become imperative to analyze the current trend and future direction of asthma cell research. This study summarized asthma literature indexed in the Medical Literature Analysis and Retrieval System Online (MEDLINE)

of the National Library of Medicine (NLM) in the past 30 years and explored the history and present state of asthma cell research by stem frequency rank to provide ideas for future work.

## 2. Objects and Methods

**2.1. Objects.** Literature of asthma cell research indexed from MEDLINE in the past 30 years was divided into three groups with 10 years as the retrieval time. The literature containing the keywords “Asthma” and “Cell” in the fields “Title” or “Abstract” was included for further investigation. The limit of publication date in the three groups were “January 1, 1987, to December 31, 1996,” “January 1, 1997, to December 31, 2006,” and “January 1, 2007, to December 31, 2016,” respectively.

The search results of each decade were exported into a CSV file with information such as title and author. All the titles of each CSV file were saved as a text file for analysis with stem frequency rank.

**2.2. Programming.** Due to the large amount data in the literature, we adopted Apache Hadoop, which is commonly used in big data analysis as the data storage framework. As a file system supporting a data-intensive distributed application, Apache Hadoop has better distribution characteristics and provides file services with both reliability and mobility for the program development [8]. To speed up the computation, we selected the Apache Spark open-source computing framework in our study instead of the Apache Hadoop built-in MapReduce computing method. The major difference between Spark and MapReduce lies in in-memory computing technology, which means the data are analyzed and processed to acquire the results in the memory before being written to the hard disk [9].

The analysis of stem frequency ranking was handled using the Natural Language Tool Kit (NLTK). NLTK is an important tool for dealing with human natural language, which can be applied to word merging, text retrieval, and statistics, and so on. The technologies such as “Word frequency Accumulation,” “Stemming Processing,” and “Stop-word Filtering” applied in this study were all performed with NLTK [10]. According to the integrated application of the above techniques, the programming environment and working process can be summarized as follows:

- (1) Programming environment: EC2 server of Amazon Web Services (AWS) platform was selected as the programming environment.

Server model: t2 micro

Server Location: Oregon, United States

Operating System and software: Ubuntu Server 14.04 with built-in Python language (version 2.7.3), Apache Spark (version 1.6.2), and NLTK (version 3.0).

- (2) Working process:

- (a) Import a text file.
- (b) Create the Spark context.

- (c) Convert all text to lowercase.
- (d) Remove punctuations, empty lines, and non-letter symbols.
- (e) Use stop word list to filter irrelevant vocabulary of research such as “they,” “where,” “to,” and “is.” The words influencing the results such as “review,” “asthma,” and “cell” were also added into the stop word list for filtering.
- (f) Stemming for reducing each word to its base form by removing its common morphological ending. In this study, we utilized PorterStemmer, a Python wrapper of the libstemmer library, to perform the stemming step.
- (g) Rank all stems according to their frequency.
- (h) List the top 50 stems, and output the results.

## 3. Results

**3.1. The Number of Studies Retrieved by Searching Three Decades.** 1331, 4393, and 7215 records were retrieved in the 1st, 2nd, and 3rd decade, respectively, which shows that the number of cell research literature of asthma indexed by MEDLINE presents explosive growth from 1987 to 2017; the literature number of the next decade was 1.5–2 times greater than the previous one.

**3.2. The Top 50 Stems of Three Decades.** The top 50 stems of 3 decades are listed in Table 1. The number of stems shared with 3 decades was 26 (Figure 1) and with the first and last 2 decades were 5 and 13 (Figures 2 and 3). The numbers of unique stems of 3 decades were 19, 6, and 11, respectively (Figure 4). According to the chronological order, the author names with the highest frequency of the three decades are Pascal Chanez (Aix-Marseille University, Paris, France), Stephen T. Holgate (University of Southampton, Southampton, United Kingdom), and Andrew Halayko (University of Manitoba, Manitoba, Canada).

## 4. Discussion

**4.1. Mainstream Research Trends in Three Decades.** The mainstream research trends can be summarized from stems shared with 3 decades. First, experimental research attracted more attention by researchers rather than clinical research. “Children” is the only relevant stem on behalf of clinical research for its frequent occurrence of asthma among children, and a study reported that asthma is common in children and is a leading cause of childhood hospitalization [11]. The stems related with experimental research, such as “respons,” “express,” and “induc,” were much more common with rising frequency year after year. Mechanism (mechan) and interventional effect were the two main directions of experimental research, but mechanism was more popular among scholars because stems about activation (activ) of cells or pathways [12–14], immune or cell response (respons) [15–17], genetic or protein expression (express) [18–20], and the role of cells or relevant genes or protein were always in the top 10 among all decades. The other 3 stems about mechanism research “product,” “induc,” and

TABLE 1: Top 50 stems of three decades.

Ranking	1st decade		2nd decade		3rd decade	
	Stem	Frequency	Stem	Frequency	Stem	Frequency
1	eosinophil	147	inflamm	545	inflamm	1153
2	activ	115	express	425	respons	552
3	respons	113	eosinophil	358	express	537
4	inflamm	102	receptor	351	t	526
5	t	90	activ	341	activ	525
6	express	84	respons	339	receptor	508
7	bronchoalveolar	80	t	326	mice	476
8	role	76	role	321	role	429
9	lavag	75	induc	298	epitheli	422
10	mast	72	mast	232	induc	410
11	lymphocyt	71	smooth	229	muscl	406
12	atop	67	muscl	228	smooth	401
13	cytokin	61	epitheli	228	inhibit	401
14	receptor	60	cytokin	211	immun	336
15	releas	59	product	207	mous	333
16	induc	58	inhibit	205	murin	330
17	inhibit	55	allergen	200	mast	328
18	allergen	54	atop	177	eosinophil	294
19	product	53	protein	171	remodel	290
20	epitheli	53	sputum	168	signal	257
21	inhal	50	gene	166	protein	250
22	adhes	49	murin	160	children	245
23	mediat	47	mice	158	cytokin	244
24	inflammatori	46	pulmonari	148	inflammatori	235
25	children	45	inflammatori	146	suppress	234
26	histamin	43	hyperrespons	144	hyperrespons	232
27	muscl	41	children	136	pulmonari	230
28	leukotrien	40	inhal	136	gene	226
29	antigen	40	th2	118	product	224
30	smooth	40	remodel	118	th2	215
31	fluid	38	inhibitor	118	modul	213
32	sodium	38	mechan	116	dendrit	210
33	ige	37	develop	116	develop	207
34	peripher	34	rat	115	allergen	196
35	macrophag	34	allergeninduc	114	attenu	195
36	pulmonari	33	immun	113	pathway	190
37	vitro	32	lymphocyt	110	differenti	190
38	mechan	32	infect	106	target	185
39	select	31	sensit	102	novel	183
40	protein	30	novel	102	infect	177
41	therapi	28	mediat	100	mediat	176
42	glucocorticoid	28	peripher	100	regulatori	174
43	bronchoconstrict	28	signal	100	type	164
44	guinea	28	leukotrien	98	sever	164
45	chang	28	allergi	97	rat	164
46	pig	28	growth	97	promot	161
47	tcell	27	chemokin	96	inhibitor	156
48	hyperrespons	27	modul	93	allergi	156
49	immun	27	mous	93	mechan	151
50	modul	27	kinas	93	potenti	150

“mediat” were mainly frequent in “production of cytokines” [21], “protein or allergen-induced” [22], and “cell-mediated” [23], respectively. Inhibition was a typical intervention effect and its stem “inhibit” was ranked in our results. For example, the following were included: the inhibition of glucocorticoids on degranulation of mast cells in allergic asthma [24], inhibition of the kinase ITK in a mouse model of asthma reduces cell death [25], and the inhibition of CD38 gene-

modified dendritic cells on murine asthma development [26].

Second, two frequent stems about pathologic changes of asthma were “inflamm” and “hyperrespons.” “Inflamm” was also in the top 10 because airway inflammation is the main expression of asthma, with mechanism research about inflammation such as etiological agents and influence factor [27, 28], being very common. Hyperresponsiveness was

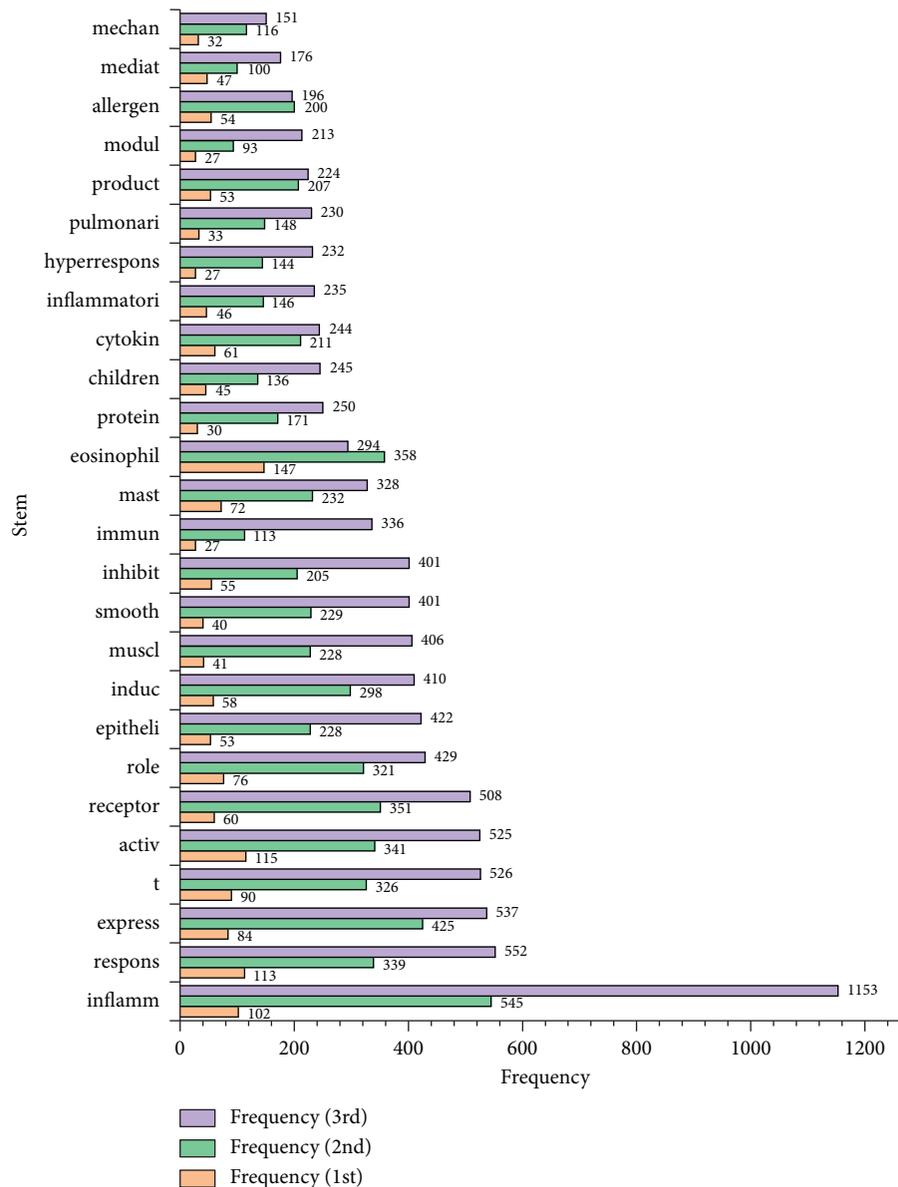


FIGURE 1: The shared stems of three decades. The orange, green, and purple bars show the frequencies of stems in the 1st, 2nd, and 3rd decade, respectively.

often mentioned with inflammation [29, 30] for immunology-related study of asthma.

Third, in terms of different types of cells related with asthma, eosinophils, mast cells, and T cells are the hot spots of immunocytes, according to the results of ranking. Mast cells are the “first line of defense” in which innate/adaptive immune cells can be activated to release a wide range of mediators by allergen-IgE-specific triggers and are widely distributed in tissues of the airway exposed to the environment, so mast cells preempt the critical roles played by histamine and mucus secretion in causing airway obstruction [31, 32]. The studies about the expression of CD antigens [33–35] involving mast cells and its mediated cytokines [36] are very common. Airway eosinophilias are associated with the inflammatory response and likely participate in airway remodeling [37–40]. Many studies have

reported that the expression of its granular proteins has functions relevant to the features of asthma, including histopathologic changes, reversible airway narrowing, and bronchial hyperreactivity [41–44]. The activated T cells in the airway wall are associated with inflammation of asthma [45, 46], and the subsets of T cell antigens have attracted extensive attention by researchers, such as the T cells of CD4+(T helper) [47–51], CD8+ [52–54], CD25+ [55], CD28 [56, 57], CD29 [58, 59], CD39+, and CD73 [60–63]. The imbalance of different subsets and the regulatory mechanism are the research emphasis of this field [64–67].

Epithelia and smooth muscle cells (SMCs) are the hot spots of structural cell studies. Research has shown that airway epithelial barrier dysfunction may have important implications for asthma [68–72]. The relevant genes or protein expression of epithelia and the regulatory

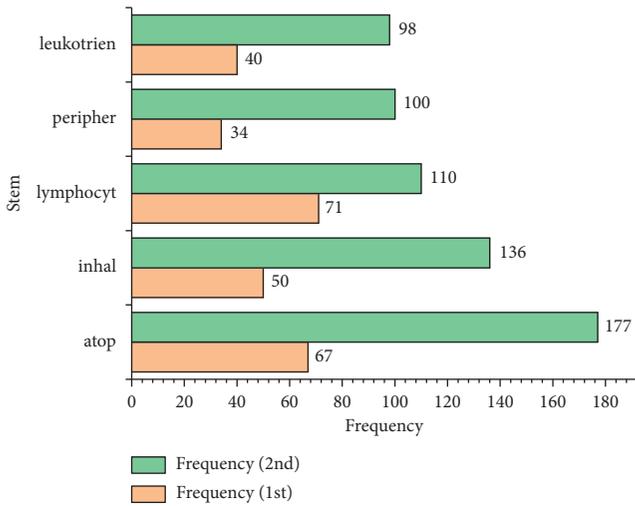


FIGURE 2: The shared stems of the first two decades. The orange and green bars show the frequencies of stems in the 1st and 2nd decade, respectively.

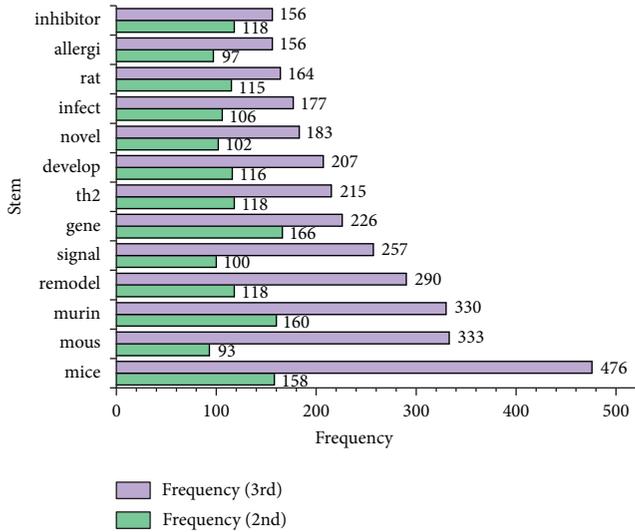


FIGURE 3: The shared stems of the last two decades. The green and purple bars show the frequencies of stems in the 2nd and 3rd decade, respectively.

mechanism of barrier function or dysfunction are the research emphasis [27, 28]. It has been reported that SMCs isolated from asthma patients release more proinflammatory mediators than that in control subjects [73], which may contribute to airway wall remodeling [74].

**4.2. Variation Trends of Research over Time.** Several variation trends can be summarized after comparing the shared stems in the first and last two decades.

First, the phenotype definition of asthma has become gradually clearer. The shared stem “atopic” in the first two decades showed that “atopic” and “non-atopic” stems were often used to define the phenotypes of asthma due to the limited available data about asthma and atopy at that time [75], which resulted in ambiguity of the phenotype

definition. However, the stem “allergi” shared with the last two decades indicated that the concept “allergic asthma” was widely used in studies [76, 77], which indicates that the phenotype of asthma was definitive.

Second, genetic studies and airway remodeling have received more attention. Along with novel experiment technologies applied in molding and detection, more studies of signaling pathways [23] and airway remodeling [78] at the gene level [18] were performed in the last 2 decades with evidence from the common stems “gene,” “signal,” and “remodel.” Moreover, the shared stems “mice,” “rat,” “murine,” and “mous” in the last two decades have shown that more animal models of rats were used in such experimental research [79–82]. In contrast, the shared stems of the first two decades implied studies of downstream signaling pathways including cytokines [83], leukotrienes [84], peripheral blood [85], or different types of receptors [86] attracted great attention at that time.

Finally, looking into changes in therapeutic approaches, the shared stem “inhal” in first two decades showed that inhaled treatment was mainstream at the early stage [87, 88]. However, it ceased to be the hot spot because of the continuous exploration of new treatments or drugs, such as Montelukast [89] and monoclonal antibodies [90]. During the past 20 years, the shared stem, “inhibitor,” indicated that as one type of a new drug for asthma, inhibitors such as histone deacetylase inhibitors [91, 92] and tyrosine kinase inhibitors [93] were implicated in influencing gene expression of asthma-related cytokines [94], gaining importance.

**4.3. Distinctive Research Hot Spots of Every Decade.**

Several distinctive research hot spots can be analyzed according to the unique stems of each decade. Two specific aspects were concerned in the studies of the first decade. First, the relevant mechanism researches including the release of cytokines [95] or histamines [95] and cell adhesion [96] about bronchoalveolar and macrophages [97, 98] were performed with corresponding experimental approaches such as cell counting method, immunofluorescence, ELISA, and bioassay commonly [99]. Second, there was specific phenomenon that researchers were enthusiastic about asthma therapy, and glucocorticoid [100] and nedocromil sodium [101] were often studied using pulmonary function test.

The main hot spot drawn from the unique stems of the second decade is that the allergen-induced topics, such as airway hyperresponsiveness [102] or inflammation and the regulation mechanism of allergic sensitization [103], were popular. Its relevant common experimental techniques included immunohistochemistry, flow cytometry, RT-PCR, and ELISA [104]. In addition, the specific research topics “different types of growth factors” [105], “kinases” [106], and “chemokines” [107] were also common due to their relevant roles that have been gradually explored and affirmed.

With the development of genetic technology, the research of the immune response became prevalent in the third decade, and specific stems about its mechanism, regulation, and signaling pathways such as “pathway” [108],

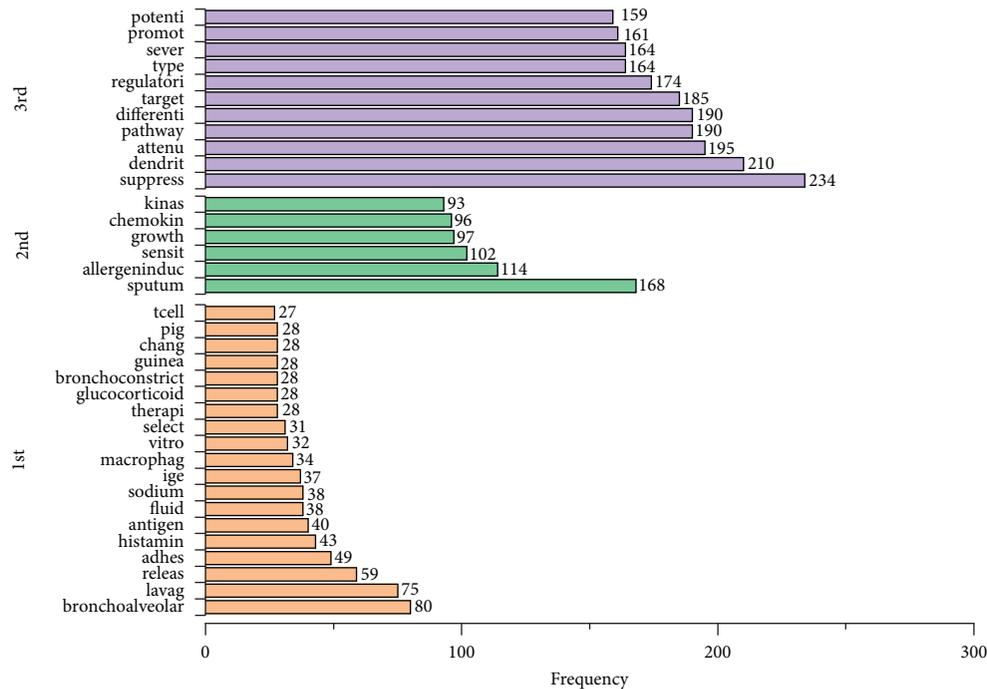


FIGURE 4: The unique stems of three decades. The orange, green, and purple bars show the frequencies of stems in the 1st, 2nd, and 3rd decade, respectively.

“target” [109], and “regulatori” [110] were found in the ranking. In terms of the corresponding experimental methods, some new techniques including digital droplet PCR (ddPCR) [111], whole-genome screen [112], and multiplexed fluorescent microsphere-based immunoassay (xMAP technology) [113] were widely adopted. Besides, a new type of cell was found in the list, dendritic cells [26], as one of the sentinel cells. Dendritic cells are the most important and primary antigen-presenting cells of asthma. They take up the antigen, process it, and present the processed antigen to T cells [114]. Therefore, dendritic cell-related studies may be one of the breakthroughs in the treatment of asthma.

## 5. Conclusion and Future Trends

The number of cell research studies of asthma indexed by MEDLINE has increased rapidly. According to the ranking list of frequent stems, scholars paid more attention to experimental research, especially mechanistic research, rather than clinical research. The immunocyte studies and structural cell research are the two main directions. Eosinophils, mast cells, and T cells are the hot spots of immunocyte studies, while epithelia and SMCs are the hot spots of structural cell research. The research trend is closely linked with the development of experimental technology, including animal models. Early studies featured basic research, but immunity research has dominated in the recent decade with the development of genetic technology.

Based on the stem rankings of three decades, future trends can be predicted in the following aspects: (1) The distinct definition of asthma phenotypes associated with

genetic characteristics will provide benefits for basic studies and clinical therapy. For instance, personalized medicine treatment tailored to individual’s asthma phenotypes identified through biomarkers [115]. (2) Immunity research involving signaling pathways, regulatory mechanisms, targets with specific biomarkers, and so on at the gene level will provide more evidence for the pathogenesis of asthma. Meanwhile, the discovery of asthma biomarkers will contribute to characterize the population and associate the disease with environmental and therapeutic effects [116], as well as predict prognosis [117]. (3) The study of new cells regulating allergy, inflammation, or remodeling of airways, such as dendritic cells, type 2 innate lymphoid cells [118], and regulatory T cells [119], will bring the potential to provide therapeutic benefits.

## Conflicts of Interest

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

## Authors’ Contributions

Yi Shang and Wenchao Tang developed and designed the study. Wenchao Tang performed the programming. Wenchao Tang wrote the paper. Bin Xiao, Peitong Wen, Ruoyun Lyu, and Ke Ning reviewed and edited the manuscript. All authors read and approved the manuscript.

## Acknowledgments

This work was supported by National Natural Science Foundation of China (Grant no. 81403469).

## References

- [1] K. Raemdonck, K. Baker, N. Dale et al., "CD4(+) and CD8(+) T cells play a central role in a HDM driven model of allergic asthma," *Respiratory Research*, vol. 17, no. 1, p. 45, 2016.
- [2] J. P. Lynch, M. A. Ferreira, and S. Phipps, "Th2/Th17 reciprocal regulation: twists and turns in the complexity of asthma phenotypes," *Annals of Translational Medicine*, vol. 4, no. S1, p. S59, 2016.
- [3] S. Koch, N. Sopel, and S. Finotto, "Th9 and other IL-9-producing cells in allergic asthma," *Seminars in Immunopathology*, vol. 39, no. 1, pp. 55–68, 2017.
- [4] S. Suzuki, M. Ogawa, S. Ohta et al., "Induction of airway allergic inflammation by hypothiocyanite via epithelial cells," *Journal of Biological Chemistry*, vol. 291, no. 53, pp. 27219–27227, 2016.
- [5] Q. M. Pei, P. Jiang, M. Yang et al., "Upregulation of a disintegrin and metalloproteinase-33 by VEGF in human airway smooth muscle cells: implications for asthma," *Cell Cycle*, vol. 15, no. 20, pp. 2819–2826, 2016.
- [6] M. Schmidt and S. Mattoli, "A mouse model for evaluating the contribution of fibrocytes and myofibroblasts to airway remodeling in allergic asthma," *Methods in Molecular Biology*, vol. 1032, pp. 235–255, 2013.
- [7] M. Mohammadian, M. H. Boskabady, I. R. Kashani et al., "Effect of bone marrow derived mesenchymal stem cells on lung pathology and inflammation in ovalbumin-induced asthma in mouse," *Iranian Journal of Basic Medical Sciences*, vol. 19, no. 1, pp. 55–63, 2016.
- [8] A. O'Driscoll, J. Daugeilaite, and R. D. Sleator, "Big data, Hadoop and cloud computing in genomics," *Journal of Biomedical Informatics*, vol. 46, no. 5, pp. 774–781.
- [9] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. U. Khan, "The rise of "big data" on cloud computing: review and open research issues," *Information Systems*, vol. 47, pp. 98–115, 2015.
- [10] T. Bekhuis, M. Kreinacke, H. Spallek, M. Song, and J. A. O'Donnell, "Using natural language processing to enable in-depth analysis of clinical messages posted to an Internet mailing list: a feasibility study," *Journal of Medical Internet Research*, vol. 13, no. 4, p. e98, 2011.
- [11] H. J. Jiang, C. A. Russo, and M. L. Barrett, *Nationwide Frequency and Costs of Potentially Preventable Hospitalizations, 2006: Statistical Brief #72*, Healthcare Cost and Utilization Project (HCUP) Statistical Briefs, Rockville, MD, USA, 2006.
- [12] I. M. de Kleer, M. Kool, M. J. de Bruijn et al., "Perinatal activation of the interleukin-33 pathway promotes type 2 immunity in the developing lung," *Immunity*, vol. 45, no. 6, pp. 1285–1298, 2016.
- [13] J. Meng, X. Xin, Z. Liu et al., "Propofol inhibits T-helper cell type-2 differentiation by inducing apoptosis via activating gamma-aminobutyric acid receptor," *Journal of Surgical Research*, vol. 206, no. 2, pp. 442–450, 2016.
- [14] S. U. Naveed, D. Clements, D. J. Jackson et al., "MMP-1 activation contributes to airway smooth muscle growth and asthma severity," *American Journal of Respiratory and Critical Care Medicine*, vol. 195, no. 8, pp. 1000–1009, 2017.
- [15] A. Kicic, P. T. Stevens, E. N. Sutanto et al., "Impaired airway epithelial cell responses from children with asthma to rhinoviral infection," *Clinical & Experimental Allergy*, vol. 46, no. 11, pp. 1441–1455, 2016.
- [16] J. W. Steinke and L. Borish, "Immune responses in rhinovirus-induced asthma exacerbations," *Current Allergy and Asthma Reports*, vol. 16, no. 11, p. 78, 2016.
- [17] L. M. Lee, M. Ji, M. Sinha et al., "Determinants of divergent adaptive immune responses after airway sensitization with ligands of toll-like receptor 5 or toll-like receptor 9," *PLoS One*, vol. 11, no. 12, Article ID e0167693, 2016.
- [18] B. D. Modena, E. R. Bleeker, W. W. Busse et al., "Gene expression correlated to severe asthma characteristics reveals heterogeneous mechanisms of severe disease," *American Journal of Respiratory and Critical Care Medicine*, vol. 195, no. 11, pp. 1449–1463, 2017.
- [19] I. Suzuki, S. Kawano, K. Komiya et al., "Inhibition of IL-13-induced periostin in airway epithelium attenuates cellular protein expression of MUC5AC," *Respirology*, vol. 22, no. 1, pp. 93–100, 2016.
- [20] J. Bigler, M. Boedigheimer, J. P. Schofield et al., "A severe asthma disease signature from gene expression profiling of peripheral blood from U-BIOPRED cohorts," *American Journal of Respiratory and Critical Care Medicine*, vol. 195, no. 10, pp. 1311–1320, 2017.
- [21] B. J. Schmiedel, G. Seumois, D. Samaniego-Castruita et al., "17q21 asthma-risk variants switch CTCF binding and regulate IL-2 production by T cells," *Nature Communications*, vol. 7, p. 13426, 2016.
- [22] M. Ahmadi, R. Rahbarghazi, M. R. Aslani, A. A. Shahbazfar, M. Kazemi, and R. Keyhanmanesh, "Bone marrow mesenchymal stem cells and their conditioned media could potentially ameliorate ovalbumin-induced asthmatic changes," *Biomedicine & Pharmacotherapy*, vol. 85, pp. 28–40, 2017.
- [23] Y. Y. Kim, I. G. Je, M. J. Kim et al., "2-Hydroxy-3-methoxybenzoic acid attenuates mast cell-mediated allergic reaction in mice via modulation of the FcepsilonRI signaling pathway," *Acta Pharmacologica Sinica*, vol. 38, no. 1, pp. 90–99, 2016.
- [24] J. Zhou, D. F. Liu, C. Liu et al., "Glucocorticoids inhibit degranulation of mast cells in allergic asthma via non-genomic mechanism," *Allergy*, vol. 63, no. 9, pp. 1177–1185, 2008.
- [25] Y. Sun, I. Peng, J. D. Webster et al., "Inhibition of the kinase ITK in a mouse model of asthma reduces cell death and fails to inhibit the inflammatory response," *Science Signaling*, vol. 8, no. 405, p. ra122, 2015.
- [26] J. Wang, W. Zhu, Y. Chen, Z. Lin, and S. Ma, "CD38 gene-modified dendritic cells inhibit murine asthma development by increasing IL-12 production and promoting Th1 cell differentiation," *Molecular Medicine Reports*, vol. 14, no. 5, pp. 4374–4382, 2016.
- [27] W. Y. Wan, F. Hollins, L. Haste et al., "NADPH oxidase-4 overexpression is associated with epithelial ciliary dysfunction in neutrophilic asthma," *Chest*, vol. 149, no. 6, pp. 1445–1459, 2016.
- [28] O. D. Solberg, E. J. Ostrin, M. I. Love et al., "Airway epithelial miRNA expression is altered in asthma," *American Journal of Respiratory and Critical Care Medicine*, vol. 186, no. 10, pp. 965–974, 2012.
- [29] J. C. Lim, F. Y. Goh, S. R. Sagineedu et al., "A semisynthetic diterpenoid lactone inhibits NF-kappaB signalling to ameliorate inflammation and airway hyperresponsiveness in a mouse asthma model," *Toxicology and Applied Pharmacology*, vol. 302, pp. 10–22, 2016.
- [30] H. Chen, X. Xu, J. Teng et al., "CXCR4 inhibitor attenuates ovalbumin-induced airway inflammation and hyperresponsiveness by inhibiting Th17 and Tc17 cell immune

- response," *Experimental and Therapeutic Medicine*, vol. 11, no. 5, pp. 1865–1870, 2016.
- [31] A. M. Gilfillan, S. J. Austin, and D. D. Metcalfe, "Mast cell biology: introduction and overview," *Advances in Experimental Medicine and Biology*, vol. 716, pp. 2–12, 2011.
- [32] A. D. Kraneveld, S. Sagar, J. Garssen, and G. Folkerts, "The two faces of mast cells in food allergy and allergic asthma: the possible concept of Yin Yang," *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, vol. 1822, no. 1, pp. 93–99, 2012.
- [33] G. U. Hong, B. S. Park, J. W. Park, S. Y. Kim, and J. Y. Ro, "IgE production in CD40/CD40L cross-talk of B and mast cells and mediator release via TGase 2 in mouse allergic asthma," *Cellular Signalling*, vol. 25, no. 6, pp. 1514–1525, 2013.
- [34] G. U. Hong, J. Y. Lim, N. G. Kim, J. H. Shin, and J. Y. Ro, "IgE and IgA produced by OX40-OX40L or CD40-CD40L interaction in B cells-mast cells re-activate FcepsilonRI or FcalphaRI on mast cells in mouse allergic asthma," *European Journal of Pharmacology*, vol. 754, pp. 199–210, 2015.
- [35] G. U. Hong, N. G. Kim, T. J. Kim, and J. Y. Ro, "CD1d expressed in mast cell surface enhances IgE production in B cells by up-regulating CD40L expression and mediator release in allergic asthma in mice," *Cellular Signalling*, vol. 26, no. 5, pp. 1105–1117, 2014.
- [36] D. Kaur, E. Gomez, C. Doe et al., "IL-33 drives airway hyper-responsiveness through IL-13-mediated mast cell: airway smooth muscle crosstalk," *Allergy*, vol. 70, no. 5, pp. 556–567, 2015.
- [37] I. D. Pavord, "Eosinophilic phenotypes of airway disease," *Annals of the American Thoracic Society*, vol. 10, pp. S143–S149, 2013.
- [38] P. Haldar, C. E. Brightling, B. Hargadon et al., "Mepolizumab and exacerbations of refractory eosinophilic asthma," *New England Journal of Medicine*, vol. 360, no. 10, pp. 973–984, 2009.
- [39] H. G. Ortega, S. W. Yancey, B. Mayer et al., "Severe eosinophilic asthma treated with mepolizumab stratified by baseline eosinophil thresholds: a secondary analysis of the DREAM and MENSA studies," *The Lancet Respiratory Medicine*, vol. 4, no. 7, pp. 549–556, 2016.
- [40] D. S. Robinson, "Mepolizumab treatment for asthma," *Expert Opinion on Biological Therapy*, vol. 13, no. 2, pp. 295–302, 2012.
- [41] A. H. Nissim Ben Efraim and F. Levi-Schaffer, "Tissue remodeling and angiogenesis in asthma: the role of the eosinophil," *Therapeutic Advances in Respiratory Disease*, vol. 2, no. 3, pp. 163–171, 2008.
- [42] S. Shimizu, H. Kouzaki, T. Ogawa, K. Takezawa, I. Tojima, and T. Shimizu, "Eosinophil-epithelial cell interactions stimulate the production of MUC5AC mucin and profibrotic cytokines involved in airway tissue remodeling," *American Journal of Rhinology & Allergy*, vol. 28, no. 2, pp. 103–109, 2014.
- [43] H. Inoue, I. Ito, A. Niimi et al., "CT-assessed large airway involvement and lung function decline in eosinophilic asthma: The association between induced sputum eosinophil differential counts and airway remodeling," *Journal of Asthma*, vol. 53, no. 9, pp. 914–921, 2016.
- [44] C. K. Kim, J. Choi, Z. Callaway, K. Iijima, G. Volcheck, and H. Kita, "Increases in airway eosinophilia and a th1 cytokine during the chronic asymptomatic phase of asthma," *Respiratory Medicine*, vol. 104, no. 10, pp. 1436–1443, 2010.
- [45] K. H. Shalaby and J. G. Martin, "Overview of asthma; the place of the Tcell," *Current Opinion in Pharmacology*, vol. 10, no. 3, pp. 218–225, 2010.
- [46] D. S. Robinson, "The role of the T cell in asthma," *Journal of Allergy and Clinical Immunology*, vol. 126, no. 6, pp. 1081–1091, 2010.
- [47] D. M. Youssef, R. M. Elbehidy, D. M. Shokry, and E. M. Elbehidy, "The influence of leptin on Th1/Th2 balance in obese children with asthma," *Jornal Brasileiro de Pneumologia*, vol. 39, no. 5, pp. 562–568, 2013.
- [48] Y. Kim, S. Lee, Y. S. Kim et al., "Regulation of Th1/Th2 cells in asthma development: a mathematical model," *Mathematical Biosciences and Engineering*, vol. 10, no. 4, pp. 1095–1133, 2013.
- [49] Q. Wu, Y. Tang, X. Hu et al., "Regulation of Th1/Th2 balance through OX40/OX40L signalling by glycyrrhizic acid in a murine model of asthma," *Respirology*, vol. 21, no. 1, pp. 102–111, 2015.
- [50] J. Wang, J. Wu, L. Kong et al., "BuShenYiQi formula strengthens Th1 response and suppresses Th2-Th17 responses in RSV-induced asthma exacerbated mice," *Journal of Ethnopharmacology*, vol. 154, no. 1, pp. 131–147, 2014.
- [51] P. Wang, D. You, J. Saravia, H. Shen, and S. A. Cormier, "Maternal exposure to combustion generated PM inhibits pulmonary Th1 maturation and concomitantly enhances postnatal asthma development in offspring," *Particle and Fibre Toxicology*, vol. 10, no. 1, p. 29, 2013.
- [52] H. Oda, T. Kawayama, H. Imaoka et al., "Interleukin-18 expression, CD8(+) T cells, and eosinophils in lungs of nonsmokers with fatal asthma," *Annals of Allergy, Asthma & Immunology*, vol. 112, no. 1, pp. 23.e1–28 e1, 2014.
- [53] A. J. Ravensberg, A. M. Slats, S. van Wetering et al., "CD8(+) T cells characterize early smoking-related airway pathology in patients with asthma," *Respiratory Medicine*, vol. 107, no. 7, pp. 959–966, 2013.
- [54] C. Ito, K. Okuyama-Dobashi, T. Miyasaka et al., "CD8+ T cells mediate female-dominant IL-4 production and airway inflammation in allergic asthma," *PLoS One*, vol. 10, no. 10, Article ID e0140808, 2015.
- [55] X. Ma, Z. A. Liang, H. Mao, Y. Liu, and M. Y. Wang, "The expression of CD25 and FOXP3 in mouse asthma and the effect of dexamethasone," *Sichuan Da Xue Xue Bao Yi Xue Ban*, vol. 41, no. 6, pp. 955–959, 2010.
- [56] K. Botturi, Y. Lacoueille, A. Cavailles, D. Vervloet, and A. Magnan, "Differences in allergen-induced T cell activation between allergic asthma and rhinitis: role of CD28, ICOS and CTLA-4," *Respiratory Research*, vol. 12, no. 1, p. 25, 2011.
- [57] M. Eusebio, L. Kraszula, M. Kupczyk, P. Kuna, and M. Pietruczuk, "The effects of interleukin-10 or TGF-beta on anti-CD3/CD28 induced activation of CD8+CD28- and CD8+CD28+ T cells in allergic asthma," *Journal of Biological Regulators and Homeostatic Agents*, vol. 27, no. 3, pp. 681–692, 2013.
- [58] S. Bazan-Socha, J. Zuk, B. Jakiela, G. Pulka, K. Pelka, and J. Musial, "Increased expression of alpha(2) (CD49b), alpha (4) (CD49d) and beta(1) (CD29) integrin subunits on peripheral blood T lymphocytes in clinically stable mild-to-moderate persistent asthma," *Polish Archives of Internal Medicine*, vol. 122, no. 12, pp. 585–590, 2012.
- [59] A. Todo-Bom, A. Mota-Pinto, V. Alves, and M. Santos-Rosa, "Aging and asthma - changes in CD45RA, CD29 and CD95 T cells subsets," *Allergologia et Immunopathologia*, vol. 40, no. 1, pp. 14–19, 2012.

- [60] P. Li, Y. Gao, J. Cao et al., "CD39+ regulatory T cells attenuate allergic airway inflammation," *Clinical & Experimental Allergy*, vol. 45, no. 6, pp. 1126–1137, 2015.
- [61] L. L. Wang, P. H. Tang, C. G. Shi et al., "Expression of CD39 mRNA is altered in the peripheral blood of patients with allergic asthma," *Biomedical Reports*, vol. 2, no. 1, pp. 75–78, 2013.
- [62] L. L. Wang, H. P. Tang, G. C. Shi et al., "CD39/CD73 and the imbalance of Th17 cells and regulatory T cells in allergic asthma," *Molecular Medicine Reports*, vol. 8, no. 5, pp. 1432–1438, 2013.
- [63] E. H. Mann, E. S. Chambers, Y. H. Chen, D. F. Richards, and C. M. Hawrylowicz, "1alpha,25-dihydroxyvitamin D3 acts via transforming growth factor-beta to up-regulate expression of immunosuppressive CD73 on human CD4+ Foxp3+ T cells," *Immunology*, vol. 146, no. 3, pp. 423–431, 2015.
- [64] N. F. Ji, Y. C. Xie, M. S. Zhang et al., "Ligustrazine corrects Th1/Th2 and Treg/Th17 imbalance in a mouse asthma model," *International Immunopharmacology*, vol. 21, no. 1, pp. 76–81, 2014.
- [65] L. Fan, X. Wang, L. Fan et al., "MicroRNA-145 influences the balance of Th1/Th2 via regulating RUNX3 in asthma patients," *Experimental Lung Research*, vol. 42, no. 8–10, pp. 417–424, 2016.
- [66] J. Wang, R. G. Jin, L. Xiao, Q. J. Wang, and T. H. Yan, "Anti-asthma effects of synthetic salidroside through regulation of Th1/Th2 balance," *Chinese Journal of Natural Medicines*, vol. 12, no. 7, pp. 500–504, 2014.
- [67] M. Zhu, Z. Liang, T. Wang, R. Chen, G. Wang, and Y. Ji, "Th1/Th2/Th17 cells imbalance in patients with asthma with and without psychological symptoms," *Allergy and Asthma Proceedings*, vol. 37, no. 2, pp. 148–156, 2016.
- [68] S. G. Royce, K. P. Patel, and C. S. Samuel, "Characterization of a novel model incorporating airway epithelial damage and related fibrosis to the pathogenesis of asthma," *Laboratory Investigation*, vol. 94, no. 12, pp. 1326–1339, 2014.
- [69] K. McLellan, M. Shields, U. Power, and S. Turner, "Primary airway epithelial cell culture and asthma in children—lessons learnt and yet to come," *Pediatric Pulmonology*, vol. 50, no. 12, pp. 1393–1405, 2015.
- [70] I. H. Heijink, M. C. Nawijn, and T. L. Hackett, "Airway epithelial barrier function regulates the pathogenesis of allergic asthma," *Clinical & Experimental Allergy*, vol. 44, no. 5, pp. 620–630, 2014.
- [71] D. A. Patel, Y. You, G. Huang et al., "Interferon response and respiratory virus control are preserved in bronchial epithelial cells in asthma," *Journal of Allergy and Clinical Immunology*, vol. 134, no. 6, pp. 1402.e7–1412 e7, 2014.
- [72] A. L. Tatler, A. E. John, L. Jolly et al., "Integrin alphavbeta5-mediated TGF-beta activation by airway smooth muscle cells in asthma," *Journal of Immunology*, vol. 187, no. 11, pp. 6094–6107, 2011.
- [73] P. Borger, N. Miglino, M. Baraket, J. L. Black, M. Tamm, and M. Roth, "Impaired translation of CCAAT/enhancer binding protein alpha mRNA in bronchial smooth muscle cells of asthmatic patients," *Journal of Allergy and Clinical Immunology*, vol. 123, no. 3, pp. 639–645, 2009.
- [74] H. Alkhouri, F. Hollins, L. M. Moir, C. E. Brightling, C. L. Armour, and J. M. Hughes, "Human lung mast cells modulate the functions of airway smooth muscle cells in asthma," *Allergy*, vol. 66, no. 9, pp. 1231–1241, 2011.
- [75] J. Pekkanen, J. Lampi, J. Genuneit, A. L. Hartikainen, and M. R. Jarvelin, "Analyzing atopic and non-atopic asthma," *European Journal of Epidemiology*, vol. 27, no. 4, pp. 281–286, 2012.
- [76] A. C. Jirno, K. Daluge, C. Happel et al., "IL-27 is essential for suppression of experimental allergic asthma by the TLR7/8 Agonist R848 (Resiquimod)," *Journal of Immunology*, vol. 197, no. 11, pp. 4219–4227, 2016.
- [77] S. S. Athari, S. M. Athari, F. Beyzay, M. Movassaghi, E. Mortaz, and M. Taghavi, "Critical role of Toll-like receptors in pathophysiology of allergic asthma," *European Journal of Pharmacology*, vol. 808, pp. 21–27, 2017.
- [78] P. Fang, H. Y. Shi, X. M. Wu et al., "Targeted inhibition of GATA-6 attenuates airway inflammation and remodeling by regulating caveolin-1 through TLR2/MyD88/NF-kappaB in murine model of asthma," *Molecular Immunology*, vol. 75, pp. 144–150, 2016.
- [79] M. E. Ferrini, S. Hong, A. Stierle et al., "CB2 receptors regulate natural killer cells that limit allergic airway inflammation in a murine model of asthma," *Allergy*, vol. 72, no. 6, pp. 937–947, 2017.
- [80] X. Diao, J. Wang, H. Zhu, and B. He, "Overexpression of programmed cell death 5 in a mouse model of ovalbumin-induced allergic asthma," *BMC Pulmonary Medicine*, vol. 16, no. 1, p. 149, 2016.
- [81] X. Y. Zhang, L. J. Ma, Y. L. Guo et al., "Effect of BCYRN1 on proliferation and migration of airway smooth muscle cells in rat model of asthma," *Zhonghua Yi Xue Za Zhi*, vol. 96, no. 46, pp. 3751–3756, 2016.
- [82] C. Tabeling, J. Herbert, A. C. Hocke et al., "Spleen tyrosine kinase inhibition blocks airway constriction and protects from Th2-induced airway inflammation and remodeling," *Allergy*, vol. 72, no. 7, pp. 1061–1072, 2017.
- [83] H. Nakajima and K. Takatsu, "Role of cytokines in allergic airway inflammation," *International Archives of Allergy and Immunology*, vol. 142, no. 4, pp. 265–273, 2007.
- [84] W. W. Busse and J. N. Gaddy, "The role of leukotriene antagonists and inhibitors in the treatment of airway disease," *American Review of Respiratory Disease*, vol. 143, no. 5, pp. S103–S107, 1991.
- [85] H. Z. Shi, S. Li, Z. F. Xie, X. J. Qin, X. Qin, and X. N. Zhong, "Regulatory CD4+CD25+ T lymphocytes in peripheral blood from patients with atopic asthma," *Clinical Immunology*, vol. 113, no. 2, pp. 172–178, 2004.
- [86] C. Duez, P. Gosset, and A. B. Tonnel, "Dendritic cells and toll-like receptors in allergy and asthma," *European Journal of Dermatology*, vol. 16, no. 1, pp. 12–16, 2006.
- [87] S. Kai, A. Nomura, Y. Morishima et al., "Effect of inhaled steroids on increased collagen synthesis in asthma," *Respiration*, vol. 74, no. 2, pp. 154–158, 2006.
- [88] K. Karwat, M. Kumor, and R. Chazan, "Coagulation parameters associated with inhaled glucocorticosteroid therapy in stable asthma," *Polski Merkuriusz Lekarski*, vol. 18, no. 107, pp. 527–529, 2005.
- [89] K. Massingham, S. Fox, and A. Smaldone, "Asthma therapy in pediatric patients: a systematic review of treatment with montelukast versus inhaled corticosteroids," *Journal of Pediatric Health Care*, vol. 28, no. 1, pp. 51–62, 2014.
- [90] M. C. Catley, J. Coote, M. Bari, and K. L. Tomlinson, "Monoclonal antibodies for the treatment of asthma," *Pharmacology & Therapeutics*, vol. 132, no. 3, pp. 333–351, 2011.
- [91] S. Toki, K. Goleniewska, S. Reiss et al., "The histone deacetylase inhibitor trichostatin A suppresses murine innate allergic inflammation by blocking group 2 innate

- lymphoid cell (ILC2) activation,” *Thorax*, vol. 71, no. 7, pp. 633–645, 2016.
- [92] J. Choi, B. K. Choi, J. S. Kim et al., “Picoside II attenuates airway inflammation by downregulating the transcription factor GATA3 and Th2-related cytokines in a mouse model of HDM-induced allergic asthma,” *PLoS One*, vol. 12, no. 1, Article ID e0167098, 2017.
- [93] A. L. da Silva, R. F. Magalhaes, V. C. Branco et al., “The tyrosine kinase inhibitor dasatinib reduces lung inflammation and remodelling in experimental allergic asthma,” *British Journal of Pharmacology*, vol. 173, no. 7, pp. 1236–1247, 2016.
- [94] P. Norman, “Inducible tyrosine kinase inhibitors: a review of the patent literature (2010–2013),” *Expert Opinion on Therapeutic Patents*, vol. 24, no. 9, pp. 979–991, 2014.
- [95] Z. L. Lummus, R. Alam, J. A. Bernstein, and D. I. Bernstein, “Characterization of histamine releasing factors in diisocyanate-induced occupational asthma,” *Toxicology*, vol. 111, no. 1–3, pp. 191–206, 1996.
- [96] I. A. Shiels, S. D. Bowler, and S. M. Taylor, “Homologous serum increases fibronectin expression and cell adhesion in airway smooth muscle cells,” *Inflammation*, vol. 20, no. 4, pp. 373–387, 1996.
- [97] T. Sano, Y. Nakamura, Y. Matsunaga et al., “FK506 and cyclosporin A inhibit granulocyte/macrophage colony-stimulating factor production by mononuclear cells in asthma,” *The European Respiratory*, vol. 8, no. 9, pp. 1473–1478, 1995.
- [98] B. Wallaert, P. Desreumaux, M. C. Copin et al., “Immunoreactivity for interleukin 3 and 5 and granulocyte/macrophage colony-stimulating factor of intestinal mucosa in bronchial asthma,” *Journal of Experimental Medicine*, vol. 182, no. 6, pp. 1897–1904, 1995.
- [99] C. Walker, E. Bode, L. Boer, T. T. Hansel, K. Blaser, and J. C. Virchow Jr., “Allergic and nonallergic asthmatics have distinct patterns of T-cell activation and cytokine production in peripheral blood and bronchoalveolar lavage,” *American Review of Respiratory Disease*, vol. 146, no. 1, pp. 109–115, 1992.
- [100] Y. Tanizaki, H. Kitani, M. Okazaki et al., “Changes in bronchoalveolar cell profiles of asthma patients induced by long-term glucocorticoid therapy,” *Nihon Kyobu Shikkan Gakkai Zasshi*, vol. 31, no. 6, pp. 717–724, 1993.
- [101] E. W. Alton and A. A. Norris, “Chloride transport and the actions of nedocromil sodium and cromolyn sodium in asthma,” *Journal of Allergy and Clinical Immunology*, vol. 98, no. 5, pp. S102–S105, 1996.
- [102] L. F. Zhou, Y. Zhu, X. F. Cui, W. P. Xie, A. H. Hu, and K. S. Yin, “Arsenic trioxide, a potent inhibitor of NF-kappaB, abrogates allergen-induced airway hyperresponsiveness and inflammation,” *Respiratory Research*, vol. 7, no. 1, 146 pages, 2006.
- [103] P. W. Hellings, A. Kasran, D. Bullens et al., “IL-10- and IL-12-independent down-regulation of allergic sensitization by stimulation of CD40 signaling,” *Journal of Immunology*, vol. 177, no. 8, pp. 5138–5144, 2006.
- [104] M. J. Leckie, A. ten Brinke, J. Khan et al., “Effects of an interleukin-5 blocking monoclonal antibody on eosinophils, airway hyper-responsiveness, and the late asthmatic response,” *The Lancet*, vol. 356, no. 9248, pp. 2144–2148, 2000.
- [105] Y. L. Tang, C. P. Hu, J. T. Feng, J. Q. Zhu, and M. J. Lin, “Regulation effect of nerve growth factor on Ras-MAPK signal transduction pathway in neurogenic inflammation of asthma,” *Zhong Nan Da Xue Xue Bao Yi Xue Ban*, vol. 31, no. 3, pp. 319–325, 2006.
- [106] W. Duan and W. S. Wong, “Targeting mitogen-activated protein kinases for asthma,” *Current Drug Targets*, vol. 7, no. 6, pp. 691–698, 2006.
- [107] D. Kaur, R. Saunders, P. Berger et al., “Airway smooth muscle and mast cell-derived CC chemokine ligand 19 mediate airway smooth muscle migration in asthma,” *American Journal of Respiratory and Critical Care Medicine*, vol. 174, no. 11, pp. 1179–1188, 2006.
- [108] Z. Wang, H. Zhang, X. Sun, and L. Ren, “The protective role of vitamin D3 in a murine model of asthma via the suppression of TGF-beta/Smad signaling and activation of the Nrf2/HO-1 pathway,” *Molecular Medicine Reports*, vol. 14, no. 3, pp. 2389–2396, 2016.
- [109] R. Zhang, H. Dong, H. Zhao, L. Zhou, F. Zou, and S. Cai, “1,25-Dihydroxyvitamin D3 targeting VEGF pathway alleviates house dust mite (HDM)-induced airway epithelial barrier dysfunction,” *Cellular Immunology*, vol. 312, pp. 15–24, 2017.
- [110] C. Martins, J. Lima, G. Nunes, and L. M. Borrego, “Regulatory T and B cells in asthmatic women: variations from pregnancy to postpartum Treg and Breg: pregnancy to postpartum,” *Journal of Investigational Allergology and Clinical Immunology*, vol. 27, no. 1, 2017.
- [111] M. Sze, Y. Oh, S. J. Yang et al., “A32 MICROBIAL INTERACTIONS AND HOST DEFENSE: The bacterial microbiome of gut, lung, and blood after H1N1 influenza virus infection in mice with and without house dust mite exposure,” *American Journal of Respiratory and Critical Care Medicine*, vol. 191, no. 1, 2015.
- [112] C. D. Campbell, K. Mohajeri, M. Malig et al., “Whole-genome sequencing of individuals from a founder population identifies candidate genes for asthma,” *PLoS One*, vol. 9, no. 8, Article ID e104396, 2014.
- [113] A. Panariti, V. Narayanan, C. Bagloli et al., “Increased vascularity of the bronchial mucosa in patients with severe asthma and the role of IL-17A in angiogenesis and vascular remodeling,” in *Proceeding of D31 Novel Mechanisms of Allergy and Airway Inflammation*, p. A6680, American Thoracic Society, New York, NY, USA, 2016.
- [114] A. S. Bharadwaj, A. K. Bewtra, and D. K. Agrawal, “Dendritic cells in allergic airway inflammation,” *Canadian Journal of Physiology and Pharmacology*, vol. 85, no. 7, pp. 686–699, 2007.
- [115] M. Godar, C. Blanchetot, H. de Haard, B. N. Lambrecht, and G. Brusselle, “Personalized medicine with biologics for severe type 2 asthma: current status and future prospects,” *MAbs*, vol. 10, no. 1, pp. 34–45, 2017.
- [116] I. Agache and L. Rogozea, “Asthma biomarkers: do they bring precision medicine closer to the clinic?,” *Allergy, Asthma & Immunology Research*, vol. 9, no. 6, pp. 466–476, 2017.
- [117] B. S. Berthon, P. G. Gibson, L. G. Wood, L. K. MacDonald-Wicks, and K. J. Baines, “A sputum gene expression signature predicts oral corticosteroid response in asthma,” *European Respiratory Journal*, vol. 49, no. 6, article 1700180, 2017.
- [118] L. Cosmi, F. Liotta, L. Maggi, and F. Annunziato, “Role of type 2 innate lymphoid cells in allergic diseases,” *Current Allergy and Asthma Reports*, vol. 17, no. 10, p. 66, 2017.
- [119] L. Bohm, J. Maxeiner, H. Meyer-Martin et al., “IL-10 and regulatory T cells cooperate in allergen-specific immunotherapy to ameliorate allergic asthma,” *Journal of Immunology*, vol. 194, no. 3, pp. 887–897, 2017.

## Research Article

# Link Prediction Investigation of Dynamic Information Flow in Epilepsy

Yan He <sup>1</sup>, Fan Yang,<sup>1</sup> Yunli Yu,<sup>2</sup> and Celso Grebogi<sup>3</sup>

<sup>1</sup>School of Biology and Engineering, Guizhou Medical University, Guiyang, Guizhou Province, China

<sup>2</sup>Department of Neurology, Affiliated Hospital of Guizhou Medical University, Guiyang, Guizhou Province, China

<sup>3</sup>Institute for Complex Systems and Mathematical Biology, King's College, University of Aberdeen, Aberdeen AB24 3UE, UK

Correspondence should be addressed to Yan He; [smileconfidence@163.com](mailto:smileconfidence@163.com)

Received 1 November 2017; Revised 3 February 2018; Accepted 19 April 2018; Published 2 July 2018

Academic Editor: Yong Xia

Copyright © 2018 Yan He 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.

As a brain disorder, epilepsy is characterized with abnormal hypersynchronous neural firings. It is known that seizures initiate and propagate in different brain regions. Long-term intracranial multichannel electroencephalography (EEG) reflects broadband ictal activity under seizure occurrence. Network-based techniques are efficient in discovering brain dynamics and offering finger-print features for specific individuals. In this study, we adopt link prediction for proposing a novel workflow aiming to quantify seizure dynamics and uncover pathological mechanisms of epilepsy. A dataset of EEG signals was enrolled that recorded from 8 patients with 3 different types of pharmacoresistant focal epilepsy. Weighted networks are obtained from phase locking value (PLV) in subband EEG oscillations. Common neighbor (CN), resource allocation (RA), Adamic-Adar (AA), and Sorenson algorithms are brought in for link prediction performance comparison. Results demonstrate that RA outperforms its rivals. Similarity matrix was produced from the RA technique performing on EEG networks later. Nodes are gathered to form sequences by selecting the ones with the highest similarity. It is demonstrated that variations are in accordance with seizure attack in node sequences of gamma band EEG oscillations. What is more, variations in node sequences monitor the total seizure journey including its initiation and termination.

## 1. Introduction

Link prediction works on revealing edge production based on network topology and node attributes. Meanwhile, higher prediction capability corresponds to more accurate description of network evolution, and it has been successfully adopted in social networks such as recommendation system. However, it is unclear that link prediction could be adopted for uncovering brain dynamics.

Brain is a complex system where multiple components work together for cognitive function. Brain is so fragile that it could be damaged forever after occasional trauma [1]. The cooperation and synergy among cerebral areas are fundamental for capability maintenance which could be investigated by graph theory as a network [2, 3]. When brain is taken as a network, individual brain areas are considered as nodes and their interaction are links. The penetration of

complex network for brain research has become popular and effective. Brain connectome reflects individual characteristics, and these wiring patterns could be taken as fingerprint of people [4]. Statistical comparison of network parameters is an objective description of brain disorders. Time-varying network changes reveal spatiotemporal alterations in the brain [5]. Certain structure affects network function, and small-worldness topology refers to quick information integration and requires neural remodeling [6]. As a cooperative element, pathological condition might stem from disconnection or rearrangement in the brain [7–9]. Investigation of pathological brain architecture changes aids in not only clinical diagnosis and early warning, but also mechanism discovery. However, more subjective description techniques are needed for investigation of evolutionary process, especially of brain progressions under pathological condition.

As a common brain disorder, epilepsy affects more than 60 million people around the world, and one third of them are refractory to medicine which require precise diagnosis and risk assessment during daily life. High amplitude synchronization is an obvious signal of epilepsy under seizure attack [10]. Multichannel EEG data records brain function under various conditions [11]. It has become indispensable in clinical diagnosis for epilepsy as an efficient mirror of neural behaviors of the brain. Detailed description of EEG signals aids in clinical evaluation and mechanism understanding. Essential component identification offers great help for understanding its working mechanism. Qualitative detection of EEG oscillations gained much achievement in aiding clinical treatment of epilepsy especially under seizure onset. Potential biomarker of epileptogenic zone includes high-frequency oscillations. Fast activity and gamma (30–100 Hz) interictal epileptiform discharges have close relationship with seizure onset according to intracranial EEG signal analysis [12, 13]. Gamma band activity used to be one significant marker for seizure occurrence [5]. Both fast beta/gamma and slow delta/theta are typical phenomena in focal epilepsies. Low frequency oscillations have been investigated as characteristic EEG patterns that influence seizure onset in temporal lobe epilepsy (TLE) [14]. Higher power and synchronization within theta band (4–7 Hz) and lower power in alpha band (10–13 Hz) demonstrate significant variations in epilepsy patients comparing with healthy controls [15].

Characterized with abnormal synchronized neural firings, the pathological mechanism of epilepsy involves whole brain network rather than certain brain area [16]. Network-based techniques are required for accounting for features of complex systems.

Multiple brain areas and routes involve in seizure course. Besides, the areas and routes vary with specific epilepsy types and patients. Spatiotemporal organization and dynamics research of EEG oscillations are necessary for precise epilepsy evaluation [17]. Basically, topographical EEG analysis quantitatively depicts seizure conditions from temporal and spatial aspect. Under certain conditions, distribution of connection probability follows the rule of power law which signifies that there exist limited brain sites connecting with most other brain areas. Meanwhile, information traffic reshapes corresponding network dynamics such as seizure attacks.

Brain connectome reveals epileptic activity by network analysis [18]. Brain network characteristics unravel functional mechanisms in clinical epilepsy [19]. Many network-based tools have been proposed to depict objective alterations during the process of epilepsy, most of them focused on evaluation indicators including degree, clustering coefficients, the shortest path length, transitivity, and efficiency. Less integrated and more detached networks are demonstrated in children with frontal lobe epilepsy (FLE) particularly when they are cognitively impaired [20]. Moreover, FLE shows more damage in regional efficiency than TLE in structural networks derived from diffusion tensor imaging [21]. Graph theory discovers brain connectivity anomalies effectively [22]. More than four subtypes are divided according to nonlinear

analysis of EEG in TLE seizures, and their medial structures demonstrate generic and disorganized network configurations [23]. Asymmetry of instant phase difference is so clear that it was capable of localizing epileptogenic zone [24]. Effects of disease duration go along with network metrics based on machine learning techniques [22]. Lower betweenness centrality reveals the effect of antiepileptic drug use [25]. Connectivity patterns detect propagation of ictal behavior during seizure onset for peeling off epilepsy [26, 27]. It is found that network structure approaches regular type while small-worldness is weakened in epilepsy [28]. In fact, it is known that small world structure produces coherent oscillations accompanied with fast response capability [29]. Although necessary, it is not easy to separate propagation area from epileptogenic zone. How to define epileptogenicity degree requires further study. What is more, not only gamma, but also delta subband EEG oscillations are assumed to be biomarkers for epilepsy, it is not clear what role frequency components play with epileptogenic zone [5]. More efforts should be put into capturing seizure network dynamics due to little consensus on epilepsy network [10].

Link prediction works on probability estimation of future edge generation based on network topology and node attributes. The capability of link prediction is in accordance with the understanding of network organization that illustrated by universal structural consistency index [30]. It is proposed that higher prediction capability corresponds to much more accurate description of network evolution [31]. Node similarity was proposed for solving link generation problem at first. Nodes sharing the same attributes tend to get connected in social networks such as recommendation system. A local naïve Bayes model uncovers hub nodes according to its description of multiple effects from their common neighbors [32]. Essential attributes are defined as node similarity when they are close to each other under certain circumstances [33, 34]. Common neighbor is the most popular structural feature that describes friends' friends. When two nodes share more common neighbors, they are easier to form new edges and the corresponding nodes would gather together into network cluster, just as what happens in social networks [35] and cooperation scientists [36]. Network cluster is an important feature of brain, and evolutionary optimization has been adopted for cluster identification of brain networks. Symmetry and matching index has been proposed for depicting cortical connectivity. The combination of network topology and energy deviation is efficient in classifying seizure occurrence in epilepsy patients [37].

In this paper, we propose a novel workflow to investigate seizure dynamics in epilepsy. Link prediction is introduced to reveal alterations in corresponding epileptogenic networks that are determined by experienced clinicians. As a preliminary work, epileptogenic networks are comprised of intrafocus sites and extrafocus sites. Since the advantage of network analysis is its fingerprint depiction of corresponding brains, nine different types of epilepsy covering 20 patients are enrolled in this study. As weight stands for information diffusion or energy cost in the brain, weighted brain networks are derived from preprocessed subband EEG

TABLE 1: Detailed information of patients.

Patient	Sex	Age	Seizure type	H/NC	Origin	Electrodes	Seizures analyzed
1	F	15	SP, CP	NC	Frontal	g, s	4
2	M	38	SP, CP, GTC	H	Temporal	D	3
3	F	26	SP, CP, GTC	H	Temporal	d, g, s	5
4	F	16	SP, CP, GTC	NC	Frontal	g, s	5
5	F	31	CP, GTC	H	Temporo/occipital	d, g, s	3
6	F	42	SP, CP, GTC	H	Temporal	D	3
7	M	47	SP, CP, GTC	H	Temporal	D	5
8	F	22	SP, CP, GTC	H	Temporo/occipital	d, s	2

oscillations by way of phase locking value. Time window is applied to generate transient dynamic frequency dependent brain networks. Four popular link prediction algorithms are adopted including CN, RA, AA, and Sorenson algorithms. The performance of link prediction capability is evaluated through AUC (area under the curve). The proposed workflow is tested on long-term EEG recordings covering seizure course. Traditional Granger causality is applied for dynamic information flow detection for comparison. Node clusters are derived by node ranking under link prediction theory. However, these node clusters are gathered together as node sequence first after node ranking. In order to distinguish the roles that played by intrafocus and extrafocus sites, transient variations of node sequences are evaluated by cosine similarity and Euclidean distance. Then node index summation is put forward to get fingerprint evaluation of each patient. The results demonstrate their efficiency in depicting seizure occurrence as well as describing epilepsy types. The layout of the paper is as follows. The materials and datasets are illustrated in detail in Section 2. The techniques and methods are presented in Section 3. Section 4 demonstrates the experimental results. Section 5 presents discussions and analyzes of the results. Finally, the conclusions are given in Section 6.

## 2. Materials

**2.1. Patient Data.** The EEG database is accessed from <https://epilepsy.uni-freiburg.de/>. A total of 8 patients with refractory focal epilepsy are enrolled in the present study. Details of the patients are illustrated in Table 1, where SP = simple partial, CP = complex partial, and GTC = generalized tonic-clonic; H = hippocampal origin and NC = neocortical origin; and d = depth electrode, g = grid electrode, and s = strip electrode. Two patients are frontal lobe epilepsy with hippocampus origin (frontal/H, short to be FLE), four patients are temporal epilepsy with hippocampus origin (temporal/H, short to be TLE), and two patients are temporal, occipital epilepsy with hippocampus origin (temporooccipital/H).

**2.2. EEG Database.** Invasive EEG recording is taken as presurgical epilepsy monitoring at the Epilepsy Center of the University Hospital of Freiburg, Germany. A Neuro-file NT digital video EEG system was applied for recording and EEG signals are sampled at 256 Hz. The recording covers the whole process of seizures. All clinical investigations have been conducted according to the principles expressed in the Declaration of Helsinki. This research was approved by the

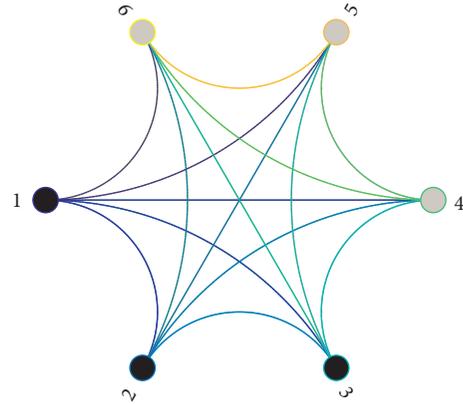


FIGURE 1: Connection among intrafocus (nodes 1–3) and extrafocus (nodes 4–6) sites in epileptogenic network.

ethics committee of University Hospital of Freiburg (Ethik-Kommission der Albert-Ludwigs-Universität Freiburg, Freiburg, Germany), and participants gave written consent for research use of these data [38–42]. Epileptic focus pervades hippocampus, neocortical area of brain structure separately, and these focuses are confirmed by experienced clinicians. These clinicians are also the guidance for the implantation of recording electrodes in the meantime, which ensures the inclusion of both intrafocus and extrafocus areas. Later, the epileptogenic focus and the accurate time for seizure occurrence are identified by experienced surgeon. Long-term EEG signals were recorded by intracranial grid, strip, and depth electrodes. For simplification, intrafocus sites are denoted as 1 to 3 and extrafocus are denoted as 4 to 6 separately as illustrated in Figure 1. It can be assumed that all nodes are connected with other ones except themselves.

Total 30 seizures in 3 epilepsy subtypes in 8 patients are investigated in this study. EEG signals are divided into preictal, ictal, and postictal sessions accordingly. No eye or muscle artifact is recorded in these intracranial EEG signals. Since the strongest noise is 50 Hz power line interference, preliminary EEG preprocessing is comprised of wavelet-based general denoising and 50 Hz power frequency interference elimination. Then signals are removed when their amplitudes exceed 8000 microvoltage. In earlier spectrogram analysis, time-varying power distribution displays in various frequency bands based on short time Fourier Transform. It is found that increased hypersynchronous neural firings occur in a large range of subband EEG oscillations and pervade many areas under seizure attack.

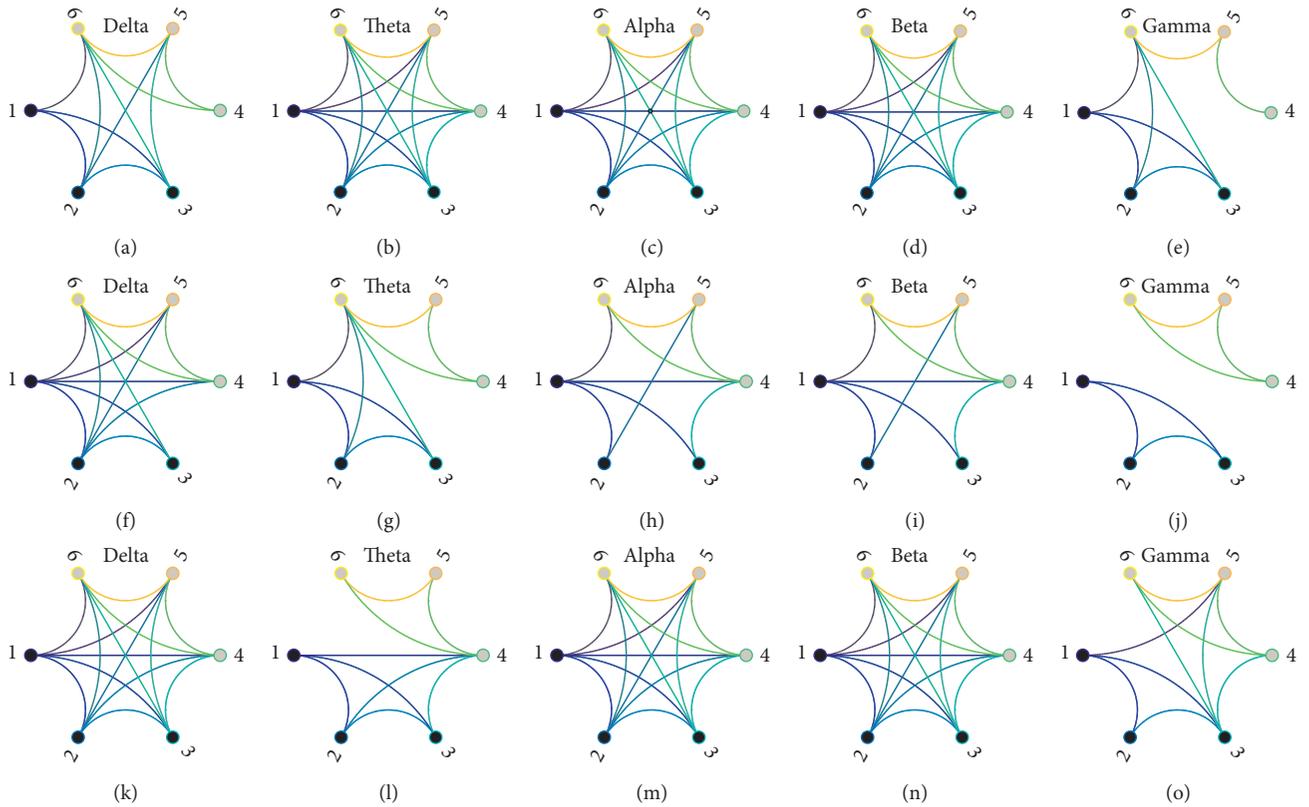


FIGURE 2: Wiring patterns during seizure course in brain network. Time-varying wiring patterns during seizure course in brain network with TLE, where weighted networks are transformed to be binary at the threshold value of 0.3. Nodes 1 to 3 denote intrafocal electrode sites while nodes 4 to 6 denote extrafocal electrode sites. (a–e) Frequency-dependent wiring links prior to seizure attack derived from delta, theta, alpha, beta, and gamma subband EEG oscillations; (f–j) frequency-dependent wiring links under seizure attack derived from delta, theta, alpha, beta, and gamma subband EEG oscillations; (k–o) frequency-dependent wiring links after seizure attack derived from delta, theta, alpha, beta, and gamma subband EEG oscillations.

**2.3. Network Dataset.** The network data for algorithm test were obtained from [www.linkprediction.org](http://www.linkprediction.org) [43], which was adopted many times by researchers in network science for detecting new techniques. It includes six datasets as follows: Jazz musician cooperation network, King James dataset, USair dataset, Adolescent dataset, NetScience dataset, and *C. elegans* dataset. Then simulated scale-free network was brought in together with real EEG network derived from epilepsy patients. The number of nodes for simulated network is 30 while its averaged node degree is set to be close to EEG network. The EEG dataset is downloaded from <http://eeg.pl/epi/> for link prediction test where the recording sites will be taken as node number while their interaction will be taken as links [44]. All datasets are divided into train set and test set for algorithm validation. The ratio for train set will not exceed 95 percent.

### 3. Methods

**3.1. Network Abstraction.** After preprocessing, EEG signals are divided into five subband oscillations as follows: delta band (0.5–4 Hz), theta band (4–8 Hz), alpha band (8–13 Hz), beta band (13–30 Hz), and gamma band (30–45 Hz). PLV is applied for detecting neural interactions among recording

sites [45], where frequency dependent time window is adopted to detect transient fluctuations. Meanwhile, non-overlapping time window is implemented for getting time variant fluctuations. Network nodes correspond with recording sites individually. All total PLV are considered as weighted connections in corresponding brain areas except that self-connections are eliminated. Higher PLV signifies stronger correlations. The weighted network is assigned to quantify frequency-dependent interaction strength. Wiring pattern is time-varying during the whole process in each subband EEG networks. Since weights stand for interaction strength and energy cost in the brain, binary networks demonstrate stronger connections as weak ties are removed. As shown in Figure 2, distinct changes occur under different situations when the threshold value for the valid link is set to be 0.3. What is more, it is roughly illustrated in Figure 2(j) that intrafocal and extrafocal sites could be separated from each other in gamma band networks when the connection strength is considered.

**3.2. Link Prediction Algorithm Comparison.** Four algorithms including CN, RA, AA, and Sorenson algorithm are introduced for link prediction capability comparison.

RA takes node degree into consideration while allocated resource depends on nodes' common neighbors [46]. In the AA technique, the nodes with smaller degree are regarded to have higher contribution. For Sorenson algorithm, the ratio between common neighbor number and its degree summation matters in link production. In order to compare the performance of these algorithms, six real networks are brought in including Jazz musician cooperation network, King James dataset, USair dataset, Adolescent dataset, NetScience dataset, and *C. elegans* dataset [43]. Simulated small-world and scale-free networks are assigned for comparison which is comprised of 30 nodes. Results show that RA technique outperforms other techniques both in real networks and simulated networks just as that reported in the previous literature [46, 47]. RA and local path index is proposed by Tao Zhou and Linyuan [33, 43]. Let  $A(V, E)$  denotes a network where  $V = \{v_1, v_2, v_3, v_n\}$  stands for node set and  $E = \{e_1, e_2, e_3, e_m\}$  stands for edge set which each edge is connected two separate nodes. Nodes and edges are important components for a network. As node degree counts the number of its neighbors, higher degree might indicate more important role it plays. CN algorithm is the most popular technique for measuring node similarity which is based on the number of common neighbors. Let  $\varphi(x)$  denotes the set of neighbors of node  $x$  and  $\varphi(y)$  denotes the neighbor set for node  $y$ , and then node similarity is defined as the number of common neighbors as follows:

$$S_{xy} = \varphi(x) \cap \varphi(y). \quad (1)$$

As illustrated in Figure 3, the similarity between node 2 and node 5 is 2, in other words,  $S_{\{2, 5\}} = 2$ .

RA considers the degree of common neighbors which act as relay station for resource transmission. The definition of RA is as follows:

$$S_{xy} = \sum_{z \in \varphi(x) \cap \varphi(y)} \frac{1}{k(z)}, \quad (2)$$

where  $\varphi(x) \cap \varphi(y)$  denotes the set of common neighbors and  $k(z)$  represents their node degree. Then  $S_{\{2, 5\}} = 1$  for nodes (2, 5) and  $S_{\{6, 3\}} = 1/4$  for nodes (3, 6) in Figure 3.

In AA algorithm, it presumes that more contribution could be achieved by the nodes that have smaller node degree. Then, its similarity is defined as follows:

$$S_{xy} = \sum_{z \in \varphi(x) \cap \varphi(y)} \frac{1}{\log k(z)}. \quad (3)$$

Then the similarity between node 2 and node 5 in Figure 3 would be  $S_{\{2, 5\}} = 1/\log 2 + 1/\log 2 = 2.8854$ .

Sorenson algorithm considers not only common neighbors, but also their degree. The definition is as follows:

$$S_{xy} = \frac{2 \times \varphi(x) \cap \varphi(y)}{k(x) + k(y)}, \quad (4)$$

$k(x)$  and  $k(y)$  denote node degree for node  $x$  and node  $y$ , respectively. Then  $S_{\{2, 5\}} = 2/3$ .

For performance comparison, the edge set  $E$  will be divided into train set  $E_t$  and test set  $E_p$ . The ratio of train set

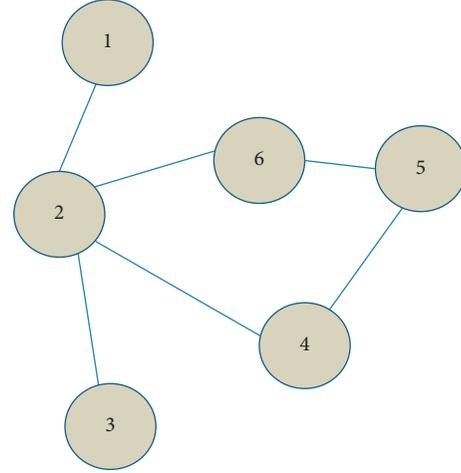


FIGURE 3: A representative network example.

will be increased from 70 percent to 95 percent, and the remaining edges will be taken as test set. Area under the receiver operating characteristic curve (AUC) is adopted for evaluating prediction accuracy. In each test, one existing link would be chosen randomly to compare with the randomly chosen nonexistent edge. Let  $n$  denotes the total test times,  $n'$  denotes the times that the higher value for existed edge than that for nonexistent edge, and  $n''$  denotes the times that two values are equal, and then AUC is defined as follows:

$$AUC = \frac{(n' + 0.5n'')}{n}. \quad (5)$$

As the value of AUC signifies the accuracy, the higher AUC indicates better performance of the corresponding algorithm. In this study, every algorithm will be tested 100 times to get the final AUC score.

According to our results, it is found out that RA outperforms other techniques as illustrated in Figure 4. In USair connection dataset, King James relationship network, Adolescent friendship network, NetScience scientists, and Jazz musician cooperation networks as well as *C. elegans* structural network, RA outperforms the other three techniques as the train ratio increased from 70 percent to 95 percent. In order to testify its efficiency in small network whose nodes not exceed 30, we brought in simulated scale free network and EEG network derived from <http://eeg.pl/epi/> [44]. It turns out that RA is better than other techniques. As illustrated in Figure 5(b), RA is also better than its candidates in brain networks, and it was adopted for further investigation.

**3.3. Node Sequence Variation Analysis and Node Index Summation.** Transient similarity matrix is derived from the RA technique by ranking nodes according to node similarity. The nodes owning highest similarity are selected as the one in the sequence correspondingly. Then, six nodes that have highest similarity are combined together as node sequence where self-similarity is eliminated. In other words, node sequences are comprised of nodes owning the largest value in similarity matrix produced from RA technique. Cosine

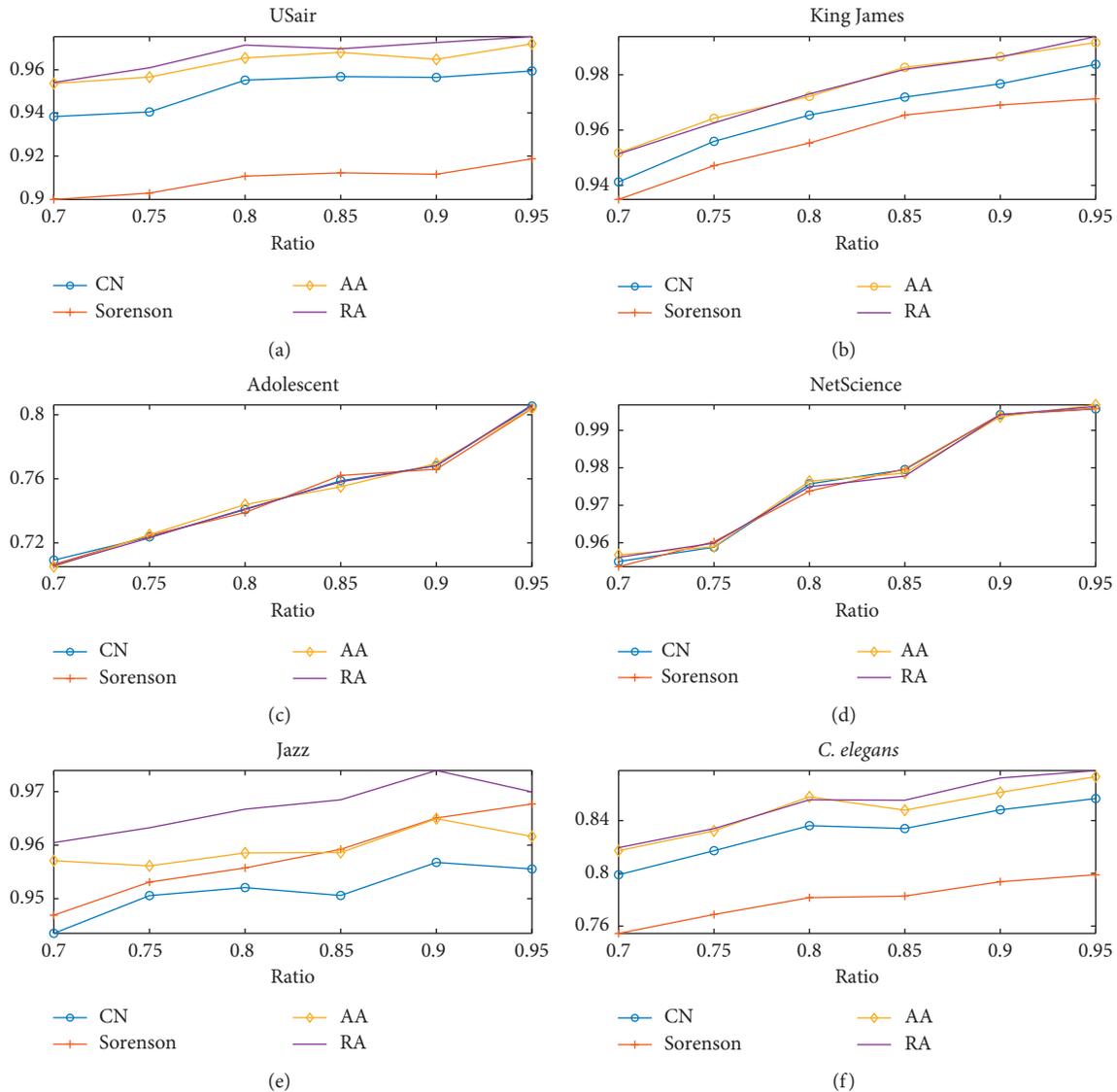


FIGURE 4: Performance comparison of link prediction algorithms on real networks: (a) USair dataset; (b) King James; (c) Adolescent; (d) NetScience; (e) Jazz; (f) *C. elegans*.

similarity and Euclidean distance are taken to measure the variation of transient node sequences, as they could measure the distance between vectors. These sequences are derived from individual adjacency matrix corresponding to preictal, ictal and postictal EEG sessions separately. Since smaller number 1 to 3 indicates intrafocus sites and higher number 4 to 6 indicates extrafocus sites, node index summation is put forward by summing up node sequences to depict the distinct roles that played by intrafocus and extrafocus points. Then node index summation is repeated in 30 seizures of 8 patients with 3 different subtypes with epilepsy. The total workflow is illustrated in Figure 6 as follows.

#### 4. Experiment and Results

Time-varying PLV was obtained by nonoverlapping time window on EEG signals. Preictal is set to be 3 minutes prior

to occurrence and postictal is set to be 3 minutes after seizure cessation. Initiation and termination of ictal process depends on clinical phenomena and was marked by experienced surgeons. According to spectrogram analysis, it is shown that higher power exist among intrafocus sites than extrafocus ones under seizure onset in TLE patients. Simultaneously, delta band oscillations dominate the energy and gamma band has minimum energy. During the seizure attack, activated nodes could be determined according to spectrogram analysis and node ranking. Besides, sharp increase and decrease of power alteration occur in gamma and delta band EEG oscillations separately.

Similarity matrix is obtained by way of RA. Node sequence is produced by combining nodes that correspond with highest value in the similarity matrix. Variation of node sequence is calculated by cosine similarity and Euclidean distance. It seems that they are sensitive to subtypes of

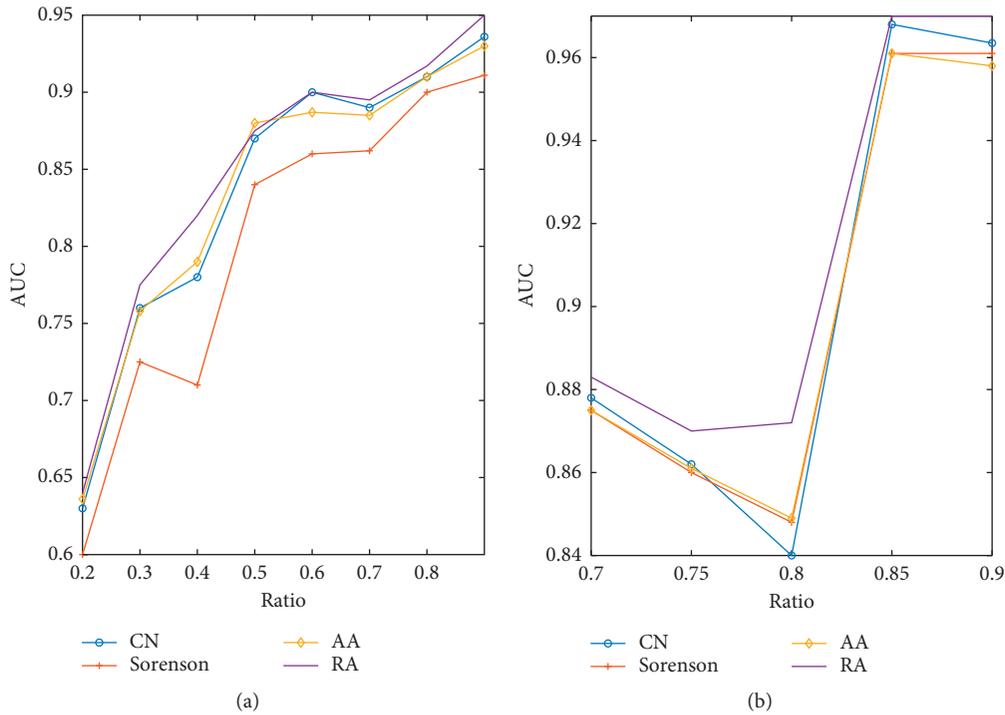


FIGURE 5: Performance comparison of four link prediction algorithms on two networks: (a) simulated scale-free network; (b) EEG network in alpha band derived from on representative epilepsy patient.

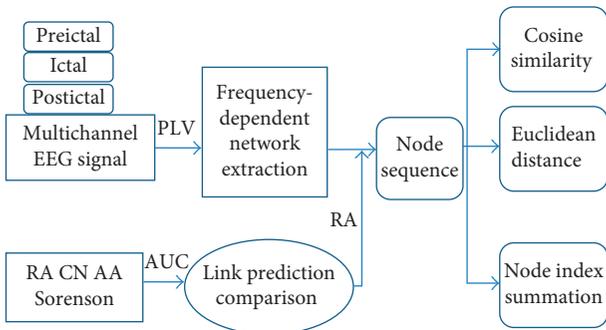


FIGURE 6: Workflow of dynamic information investigation based on link prediction algorithm.

epilepsy. As illustrated in Figure 7, from a patient with TLE, fluctuations in node sequences are both time-dependent and frequency-dependent in all EEG networks. It is demonstrated that network organization alters sharply under seizure condition especially in patients with TLE. The variation of node sequences changes strongly compared with that under preictal and postictal condition. Besides, intense network variations are demonstrated both in ictal to preictal and in ictal to postictal comparisons. Specifically, alterations of node sequences could detect seizure onset from gamma band EEG oscillations, while no clear difference are revealed in the other four subband EEG oscillations. The variation trend is in the opposite direction between temporo/occipital and TLE subtypes, which decrease at first and bounce back later during ictal session in the former, and the variation

decreases during seizure occurrence with an increase occurs prior to attack in the latter. In temporooccipital/NC subtypes, neither clear decrease nor increase springs out. Figures 8 and 9 display distinct variations of network connections in gamma band in different epilepsy subtypes. Specifically, this time-dependent variation occurs in both temporo/occipital and TLE in descending and ascending directions during the attack process separately.

Specifically, distance of ranked node sequence reflects the time-varying network organizations during seizure process. Besides, the contrary alteration of node sequence variations indicates distinct pathological mechanism of seizures in TLE and temporo/occipital epilepsy. All these attest that link prediction techniques enabling uncovering heterogeneous mechanism of epilepsy, and they might provide new insights for epilepsy investigation.

In order to separate intrafocus from extrafocus sites, nodes are ranked by their similarity according to link prediction analysis. Then, node index summation is presented for characterizing time-varying networks for seizures after node ranking in similarity matrix. Node index summation is calculated in total 8 patients over five subbands. Results demonstrate that under seizure attack condition, the networks tend to be closer to intrafocus. Since 1 to 3 indicates intrafocus sites, higher values of node index summation demonstrate more extrafocus involving in network sequence. As for FLE, more extrafocus sites take part in network sequence in delta band comparing with preictal and postictal conditions. However, more intrafocus focus sites are involved in the other four subbands. In TLE, the roles of

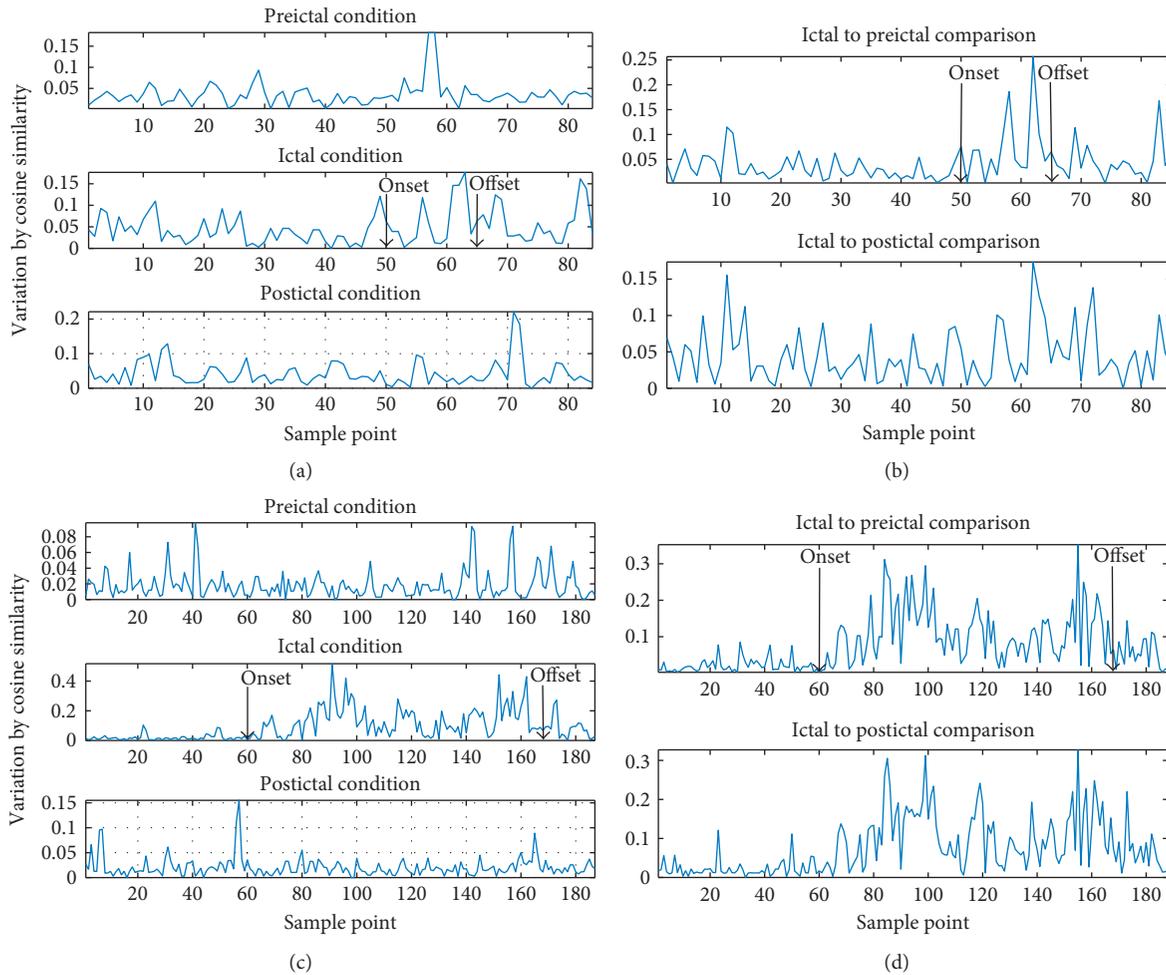


FIGURE 7: Distance of node sequences during seizure course: (a) alterations vary with time during preictal, ictal, and postictal segments in gamma band EEG networks from one patient with FLE; (b) distance of node sequences when comparing preictal to ictal and ictal to postictal segments in gamma band EEG networks from one patient with FLE; (c) alterations vary with time during preictal, ictal, and postictal segments in gamma band EEG networks from one patient with temporo/occipital epilepsy; (d) distance of node sequences when comparing preictal to ictal and ictal to postictal segments in gamma band EEG networks from one patient with temporo/occipital epilepsy.

intrafocus and extrafocus sites are different as more intrafocus takes charge during seizure process in alpha, beta, and gamma subbands. In temporooccipital epilepsy, intrafocus sites are dominant under ictal condition in alpha, beta, and gamma subbands, while extrafocus sites take charge under seizure attack in delta band, and theta band EEG networks involve more extrafocus sites after seizure termination. As node index summation indicates different roles played by intrafocus and extrafocus sites, the higher value corresponds with more extrafocus sites dominate in the network. It is illustrated that most intrafocus sites take part in network sequence under seizure occurrence. What is more, such tendency is frequency dependent. In gamma band, the networks are more obvious for reflecting seizure occurrence than other subband EEG networks under pathological condition.

We also calculated network metrics on time-varying weighted networks including weighted degree, weighted clustering coefficients, and weighted path length. These measurements could detect the eruption of seizures in the

beginning. However, it fails to detect the alteration of network organization during seizure course.

What is more, we have tested these link prediction techniques on other biological networks including simulated hierarchical and modular networks. It has been found out that RA technique is not applicable for these networks, which might due to their special topology property Figure 10.

## 5. Discussion

Combining link prediction technique with theoretical network topology, we propose a novel workflow for describing and quantifying seizures from intracranial EEG recordings. At first, EEG oscillations are divided into delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz), and gamma (30–45 Hz) bands separately. A frequency-dependent length of window is brought in to detect transient connection behavior by multichannel EEG signals, whereas time-varying PLVs are derived from these subband EEG oscillations. These PLVs are taken as elements in adjacency

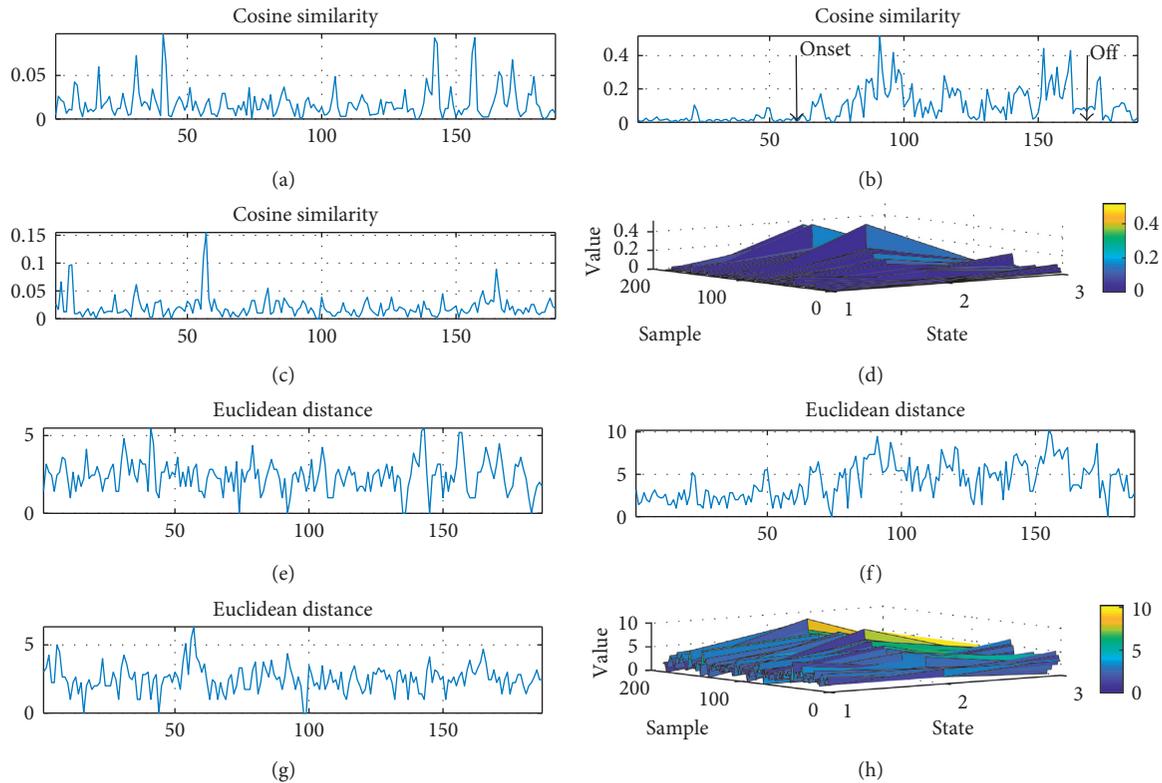


FIGURE 8: Variation of node sequences during seizure course from a patient with temporo/occipital epilepsy: (a–c) variations detected by cosine similarity under preictal, ictal, and postictal condition separately; (d) variations in preictal, ictal, and postictal condition detected by cosine similarity together; (e–g) variations detected by Euclidean distance under preictal, ictal, and postictal condition separately; (h) variations in preictal, ictal, and postictal condition detected by Euclidean distance together.

matrix. Simultaneously, CN, RA, AA, and Sorenson algorithm are brought in for performance comparison in link prediction based on node similarity. As RA defeats its opponents, the RA technique is adopted to get largest similarity matrix from the weighted networks. Node sequences are detected through these similarity matrixes during seizure condition in nine different kinds of epilepsy. Cosine similarity and Euclidean distance are applied for calculating the variation of node sequence during the total seizure process. Node index summation is put forward for separating intrafocus and extrafocus sites. It is demonstrated that variation in node sequences detect seizure bursts and monitor seizure journey in focal and temporal epilepsy. A majority of networks bounce back to previous condition in these subband EEG oscillations. This might be a reflection of brain plasticity under pathological condition. Take gamma band EEG oscillations as an example; the alteration of wiring patterns increased in temporo/occipital epilepsy with hippocampus origin under seizure onset, while the alterations decreased in TLE with hippocampus origin under the same circumstances. According to node index summation, different roles are played by intrafocus and extrafocus sites during seizure journey. In other words, the involvement is changing of intrafocus and extrafocus sites in the epileptogenic network. What is more, patients' specific information could be obtained by node index summation and frequency dependency appears on epilepsy subtypes.

Normally, social networks are different from biological networks in assortative mixing which makes them easy to percolate and hard to be attacked on high-degree vertex [48]. The powerful capability of link prediction has been testified in social networks. However, the adoption of link prediction is rare in brain networks. Apart from precise localization of the epileptogenic zone, it is important to describe the dynamics in the related cortical area [16–18]. If recording sites are taken as nodes, and their relations are regarded as links, nodes and edges are fundamental elements of the network. Interactions among cortical regions could be illustrated through weighted complex networks. Recurrent seizure occurrence affects brain structure and its function. Tracking seizure dynamics could unravel functional integration and segregation and then aid in clinical treatment. With the advent of complex network, EEG recordings could be investigated for uncovering working mechanisms of the brain integrally and systematically, which reflects faster temporal spontaneous neural oscillations in the brain. Network identification and characterization aids in clinical treatment for medically intractable epilepsy patients [27]. Epileptogenic networks consist focus and other cortical areas involved in seizure initiation and propagation.

In this study, the implantation of recording electrodes covers both intrafocus and extrafocus brain areas. Nodes 1 to 3 denote intrafocus area and nodes 4 to 6 correspond with extrafocus area. Long-term invasive EEG signals are

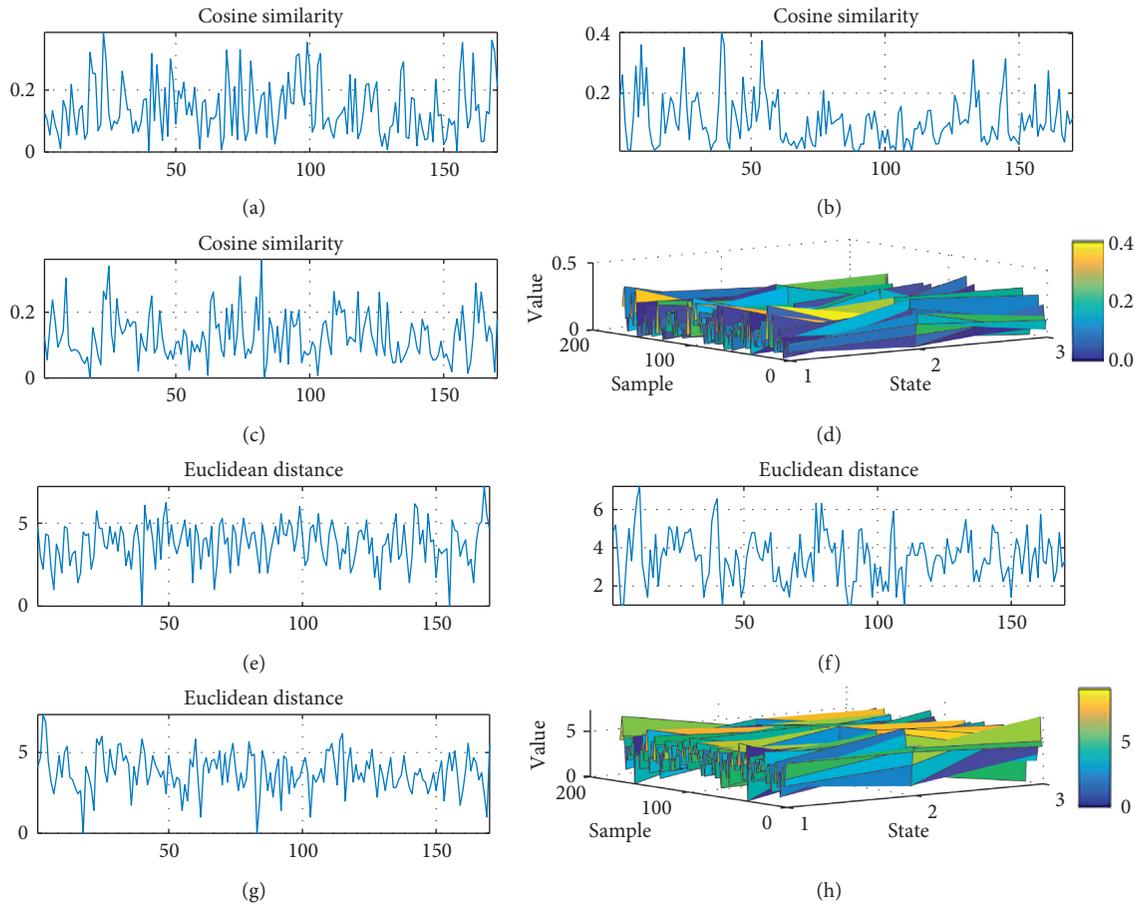


FIGURE 9: Variation of node sequences during seizure course from a patient with TLE: (a–c) variations detected by cosine similarity under preictal, ictal, and postictal condition separately; (d) variations in preictal, ictal, and postictal condition detected by cosine similarity together; (e–g) variations detected by Euclidean distance under preictal, ictal, and postictal condition separately; (h) variations in preictal, ictal, and postictal condition detected by Euclidean distance together distance of node sequences during seizure course.

recorded prior to surgery in refractory epilepsy patients. Abnormal neural firings could be eliminated or weakened according to disruption or control of these epileptogenic networks. Previously, experimental results demonstrate that link prediction on weighted network is worse than those on undirected binary network [33]. However, binary links would remove the important message including cooperation strength and energy cost. What is more, such worse results do not appear in our experiment.

Connections among brain regions quantify information flow. Edge weights reflect the similarity distance among brain areas. It is assumed that weak links play important roles in psychiatric pathologies as well as in network function [33]. In other words, weight should be considered as an important component in brain networks. In this work, our real networks contain both binary and weighted ones, and RA technique outperforms other three methods in link prediction. The power of RA technique reflects their general topology property along with network dynamics in these tested networks. Node similarity is an effective indicator for investigating network dynamics [49, 50]. All the adopted four techniques are focused on node similarity. Results demonstrate that this node similarity-based prediction

technique could detect seizure attack according to investigating weighted frequency dependent networks.

It is known that brain has to work cooperatively and harmoniously. Disconnected network in brain disorder was proposed by Wernicke in the late 19th century. However, most work focused on quantitative measurement instead of uncovering details. What is more, the small-worldness assumption of brain network requires further investigation [51]. In this paper, we brought in link prediction techniques for describing and quantifying seizures from intracranial EEG recordings. The sliding window enables capturing frequency-dependent network evolution during the course of seizures [52]. Intrafocus and extrafocus sites are denoted as 1 to 3 and 4 to 6 individually. These nodes are ranked to form node sequence according to the largest similarity matrix which was evaluated by RA technique. Nodes that have highest similarity would be combined to form node sequences corresponding to individual recording sites. As an example, the sequence might be 6,54,211 during preictal period and then it changes to 2,21,111 under seizure attack. Cosine similarity is a measure for quantifying distance between vectors and string in machine learning and artificial intelligence. The cosine similarity among these varying node

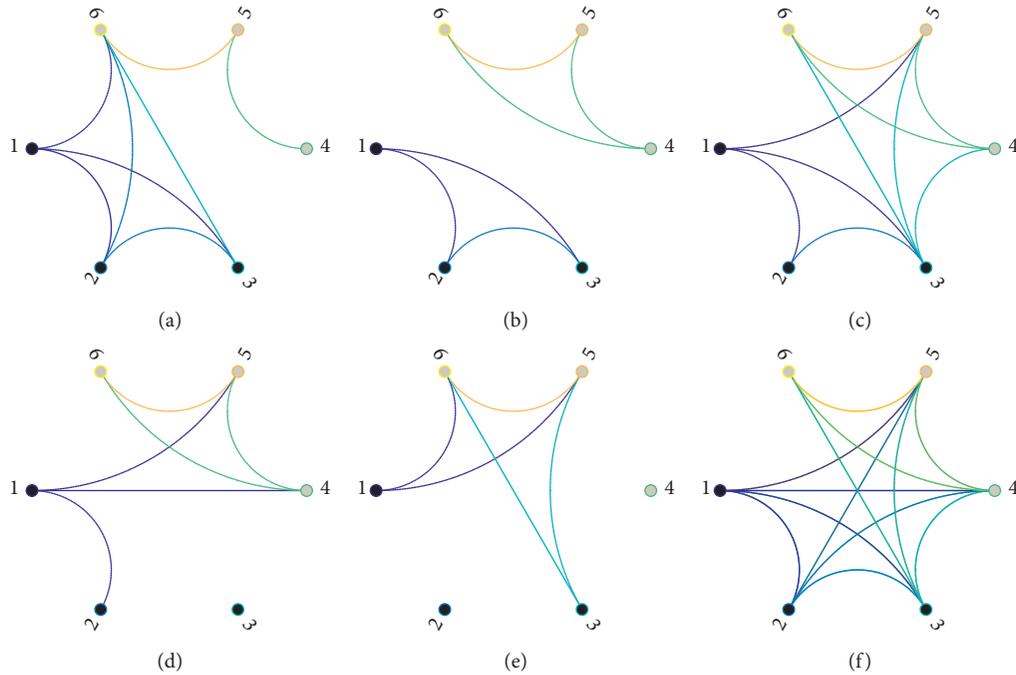


FIGURE 10: Time-varying wiring patterns in gamma band EEG oscillations: (a–c) binary brain networks for preictal, ictal, and postictal stages at the threshold value of 0.3 in patient with TLE; (d–f) binary brain networks for preictal, ictal, and postictal stages at the threshold value of 0.3 in patient with temporooccipital/H epilepsy.

sequences depicts the process of seizures especially in gamma band EEG networks.

It is illustrated that variation of node sequence detects seizure occurrence and monitors seizure process in weighted networks derived from gamma band oscillation. Take TLE as an example; alteration of network organization bursts as soon as seizure occurs in gamma band EEG oscillation, and network disruption is demonstrated in other subband EEG oscillations under attack. Our proposed workflow offers a technique for quantifying network alterations effectively. These various alterations implicit heterogeneity of network evolution accompanied with transient ictal behaviors in distinct epilepsy.

According to our research, RA has the best performance in link prediction based on node similarity. It is assumed that the consideration of the next nearest neighbors enables it to discover transmission and connectivity capability under nonlinear condition [33, 46, 53]. Besides, allocated resource might be better for representing weighted networks. One node's resource could be distributed and flowed to its neighbors, which was determined by network topology, whereas power-law strength-degree correlation produces equilibrium in resource-allocation dynamics [53]. It is postulated that RA has splendid performance in networks owning high clustering coefficients. Brain is such a high-efficient complex system that higher clustering coefficients and longer path length might coexist under pathological epileptic condition. This might contribute to the good performance of the RA technique. As RA has clear physical

significance, we could presume that these frequency-dependent varying node sequences and their distances are another reflection of network dynamics in seizures.

## 6. Conclusion

Epileptic seizure investigation from multichannel EEG signals is critical for uncovering the pathology of epilepsy. In this paper, we propose a novel workflow for depicting seizure occurrence in refractory focal epilepsy. This network-based workflow detects the seizure occurrence and monitors the total seizure course, which is the first work aiming at monitoring alteration of network reorganizations throughout seizures as far as we know. The key point is that link prediction technique is adopted to describe connection transitions under seizure attack by way of similarity matrix. It finds out that network variation detecting outburst as well as termination of seizures according to gamma band EEG oscillations in frontal and temporal focal epilepsy.

Weighted dynamic network organization reveals information transmission and interaction strength during seizure course. Results demonstrate that our workflow takes the picture of network connection transitions all the way to seizure when compared with ordinary network metrics including weighted clustering coefficients. In the beginning, four link prediction techniques are brought in for performance comparison including CN, RA, AA, and Sorenson algorithm which have demonstrated their powerful capability in social networks based on evaluating local property

of networks. It is supposed that higher predictability corresponds with better description of network dynamics. Since RA is superior to other three rivals, node sequences are produced from similarity matrix by ranking nodes based on RA technique. Node sequence variations are calculated covering seizure onset in subband EEG networks in all subtypes. Cosine similarity is better than Euclidean distance in detecting variance in these node sequences during the whole seizure process. Experimental results reveal that it could quantitatively characterize and monitor seizure occurrence. Distinct variations in node sequences are demonstrated in various epilepsy subtypes. These variations might be a reflection of their heterogeneous pathology. Node index summation quantifies both temporal and frequency information about subtype epilepsy. In conclusion, link prediction might be an efficient assistant technique for uncovering brain network dynamics in epilepsy.

The presented workflow introduces link prediction technique for depicting brain networks. Epileptogenic cortical network plays dominant roles in working mechanism of epilepsy. Heterogeneity is a critical problem when investigating epilepsy's pathology. The combination with link prediction and network organization offers new insights for understanding epilepsy through the point of seizure occurrence. In refractory focal epilepsy, surgery is a good choice to ease the pain of patients. And the separation of intrafocus and extrafocus sites aids in diagnosis and clinical treatment in epilepsy research. According to our experimental results, RA ranks individual nodes efficiently while variation in node sequences detects seizure occurrence. Limited by the number of subjects in this research, our preliminary work did not get statistical results for characterizing distinct epileptic brain networks. Therefore, more research is needed for deeper understanding. For example, more investigation should be taken to uncover the contrary changes between TLE and temporo/occipital epilepsy brain especially in gamma band EEG oscillations. What is more, RA should be investigated further by computational neuroscience methods. Besides, as gamma band EEG oscillation is believed to play important role in higher cognitive function; distinguishable alteration in gamma band networks needs more attention in epilepsy research.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Acknowledgments

This work was supported partly by the National Natural Science Foundation of China (Grant nos. 81460206 and 81660298), Scientific Research Foundation for Doctors of Guizhou Medical University (no. Yuan Bo He J [2014] 003) and the 2011 Collaborative Innovation Program of Guizhou Province (no. 2015-04 to Zhu Zeng, the Principal Investigator).

## References

- [1] S. M. Lai, S. Studenski, P. W. Duncan, and S. Perera, "Persisting consequences of stroke measured by the stroke impact scale," *Stroke*, vol. 33, no. 7, pp. 1840–1844, 2002.
- [2] R. E. Passingham, K. E. Stephan, and R. Kötter, "The anatomical basis of functional localization in the cortex," *Nature Reviews Neuroscience*, vol. 3, no. 8, pp. 606–616, 2002.
- [3] O. Sporns, G. Tononi, and R. Kötter, "The human connectome: a structural description of the human brain," *PLoS Computational Biology*, vol. 1, no. 4, pp. 245–251, 2005.
- [4] T. Kaufmann, D. Alnæs, N. T. Doan, C. L. Brandt, O. A. Andreassen, and L. T. Westlye, "Delayed stabilization and individualization in connectome development are related to psychiatric disorders," *Nature Neuroscience*, vol. 20, no. 4, pp. 513–515, 2017.
- [5] O. Grinenko, J. Li, J. C. Mosher et al., "A fingerprint of the epileptogenic zone in human epilepsies," *Brain*, vol. 141, no. 1, pp. 117–131, 2018.
- [6] R. Ferri, F. Rundo, O. Bruni, M. G. Terzano, and C. J. Stam, "Small-world network organization of functional connectivity of EEG slow-wave activity during sleep," *Clinical Neurophysiology*, vol. 118, no. 2, pp. 449–456, 2007.
- [7] E. van Diessen, S. J. Diederer, K. P. Braun, F. E. Jansen, and C. J. Stam, "Functional and structural brain networks in epilepsy: what have we learned?," *Epilepsia*, vol. 54, no. 11, pp. 1855–1865, 2013.
- [8] E. Bagarinao, S. Maesawa, Y. Ito et al., "Detecting sub-second changes in brain activation patterns during interictal epileptic spike using simultaneous EEG-fMRI," *Clinical Neurophysiology*, vol. 129, no. 2, pp. 377–389, 2017.
- [9] T. M. Medvedeva, M. V. Sysoeva, G. van Luijckelaar, and I. V. Sysoev, "Modeling spike-wave discharges by a complex network of neuronal oscillators," *Neural Networks*, vol. 98, pp. 271–282, 2017.
- [10] M. P. Richardson, "New observations may inform seizure models: very fast and very slow oscillations," *Progress in Biophysics and Molecular Biology*, vol. 105, no. 1-2, pp. 5–13, 2011.
- [11] W.-L. Zhang and B.-L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Transactions on Autonomous Mental Development*, vol. 7, no. 3, pp. 162–175, 2015.
- [12] R. G. Andrzejak, O. David, V. Gnatkovsky et al., "Localization of epileptogenic zone on pre-surgical intracranial eeg recordings: toward a validation of quantitative signal analysis approaches," *Brain Topography*, vol. 28, no. 6, pp. 832–837, 2015.
- [13] L. Ren, M. T. Kucewicz, J. Cimbalko et al., "Gamma oscillations precede interictal epileptiform spikes in the seizure onset zone," *Neurology*, vol. 84, no. 6, pp. 602–608, 2015.
- [14] J. X. Tao, X.-J. Chen, M. Baldwin et al., "Interictal regional delta slowing is an EEG marker of epileptic network in temporal lobe epilepsy," *Epilepsia*, vol. 52, no. 3, pp. 467–476, 2011.
- [15] M. A. Quraan, C. McCormick, M. Cohn, T. A. Valiante, and M. P. McAndrews, "Altered resting state brain dynamics in temporal lobe epilepsy can be observed in spectral power, functional connectivity and graph theory metrics," *PLoS One*, vol. 8, Article ID e68609, 2013.
- [16] M. A. Lopes, M. P. Richardson, E. Abela et al., "An optimal strategy for epilepsy surgery: disruption of the rich-club?," *PLoS Computational Biology*, vol. 13, no. 8, p. e1005637, 2017.
- [17] P. Jin, D. Wu, X. Li, L. Ren, and Y. Wang, "Towards precision medicine in epilepsy surgery," *Annals of Translational Medicine*, vol. 4, no. 2, p. 24, 2016.
- [18] H. Stefan and F. H. Lopes da Silva, "Epileptic neuronal networks: methods of identification and clinical relevance," *Frontiers in Neurology*, vol. 4, p. 8, 2013.

- [19] H. Burianová, N. L. Faizo, M. Gray, J. Hocking, G. Galloway, and D. Reutens, "Altered functional connectivity in mesial temporal lobe epilepsy," *Epilepsy Research*, vol. 137, pp. 45–52, 2017.
- [20] M. J. Vaessen, J. F. Jansen, M. C. Vlooswijk et al., "White matter network abnormalities are associated with cognitive decline in chronic epilepsy," *Cerebral Cortex*, vol. 22, no. 9, pp. 2139–2147, 2012.
- [21] E. Widjaja, M. Zamyadi, C. Raybaud, O. C. Snead, S. M. Doesburg, and M. L. Smith, "Disrupted global and regional structural networks and subnetworks in children with localization-related epilepsy," *American Journal of Neuroradiology*, vol. 36, no. 7, pp. 1362–1368, 2015.
- [22] M. J. Paldino, W. Zhang, Z. D. Chu, and F. Golriz, "Metrics of brain network architecture capture the impact of disease in children with epilepsy," *NeuroImage: Clinical*, vol. 13, pp. 201–208, 2017.
- [23] F. Bartolomeia, F. Wendling, J.-J. Bellanger, J. ReÂgis, and P. Chauvel, "Neural networks involving the medial temporal structures in temporal lobe epilepsy," *Clinical Neurophysiology*, vol. 112, no. 9, pp. 1746–1760, 2001.
- [24] G. J. Ortega, R. G. Sola, and J. Pastor, "Complex network analysis of human ECoG data," *Neuroscience Letters*, vol. 447, no. 2–3, pp. 129–133, 2008.
- [25] Z. Haneef, H. S. Levin, and S. Chiang, "Brain graph topology changes associated with anti-epileptic drug use," *Brain Connectivity*, vol. 5, no. 5, pp. 284–291, 2015.
- [26] P. Neal Taylor, M. Kaiser, and J. Dauwels, "Structural connectivity based whole brain modelling in epilepsy," *Journal of Neuroscience Methods*, vol. 236, pp. 51–57, 2014.
- [27] C. Wilke, G. Worrell, and B. He, "Graph analysis of epileptogenic networks in human partial epilepsy," *Epilepsia*, vol. 52, no. 1, pp. 84–93, 2011.
- [28] M. J. Vaessen, J. F. Jansen, P. A. Hofman et al., "Impaired small-world efficiency in structural brain networks in chronic epilepsy," *NeuroImage*, vol. 47, p. S113, 2009.
- [29] L. F. Lago-Fernandez, R. Huerta, F. Corbacho, and J. A. Siguenza, "Fast response and temporal coherent oscillations in small-world networks," *Physical Review Letters*, vol. 84, no. 12, pp. 2758–2761, 2000.
- [30] L. Lü, L. Pan, T. Zhou, Y.-C. Zhang, and H. Eugene Stanley, "Toward link predictability of complex networks," *Proceedings of the National Academy of Sciences*, vol. 112, no. 8, pp. 2325–2330, 2015.
- [31] Q.-M. Zhang, X.-K. Xu, Y.-X. Zhu, and T. Zhou, "Measuring multiple evolution mechanisms of complex networks," *Scientific Reports*, vol. 5, no. 1, p. 10350, 2015.
- [32] Z. Liu, Q. M. Zhang, L. Lü, and T. Zhou, "Link prediction in complex networks: a local naïve Bayes model," *EPL*, vol. 96, no. 4, p. 48007, 2011.
- [33] L. Lü and T. Zhou, "Link prediction in weighted networks: the role of weak ties," *EPL*, vol. 89, no. 1, pp. 1–6, 2010.
- [34] L. Getoor and C. P. Diehl, "Link mining: a survey," *ACM SIGKDD Explorations Newsletter*, vol. 7, no. 2, pp. 3–12, 2005.
- [35] G. Kossinets, "Effects of missing data in social networks," *Social Networks*, vol. 28, no. 3, pp. 247–268, 2006.
- [36] M. E. J. Newman, "Clustering and preferential attachment in growing networks," *Physical Review E*, vol. 64, no. 2, p. 025102, 2001.
- [37] Z. K. Gao, Q. Cai, Y. X. Yang, N. Dong, and S. S. Zhang, "Visibility graph from adaptive optimal kernel time-frequency representation for classification of epileptiform EEG," *International Journal of Neural Systems*, vol. 27, no. 4, p. 1750005, 2017.
- [38] C. Alvarado-Rojas, M. Valderrama, A. Fouad-Ahmed, H. Feldwisch Drentrup, M. Ihle, and C. A. Teixeira, "Slow modulations of high-frequency activity (40–140 Hz) discriminate preictal changes in human focal epilepsy," *Scientific Reports*, vol. 4, no. 1, p. 4545, 2015.
- [39] B. Schelter, H. Feldwisch-Drentrup, J. Timmer, J. Gotman, and A. Schulze-Bonhage, "A common strategy and database to compare the performance of seizure prediction algorithms," *Epilepsy and Behavior*, vol. 17, no. 2, pp. 154–156, 2010.
- [40] T. Maiwald, M. Winterhalder, R. Aschenbrenner-Scheibe, H. U. Voss, A. Schulze-Bonhage, and J. Timmer, "Comparison of three nonlinear seizure prediction methods by means of the seizure prediction characteristic," *Physica D: Nonlinear Phenomena*, vol. 194, no. 3–4, pp. 357–368, 2004.
- [41] R. Aschenbrenner-Scheibe, T. Maiwald, M. Winterhalder et al., "How well can epileptic seizures be predicted? An evaluation of a nonlinear method," *Brain*, vol. 126, no. 12, pp. 2616–2626, 2003.
- [42] A. Schad, K. Schindler, M. Winterhalder et al., "Application of a multivariate seizure detection and prediction method to non-invasive and intracranial long-term EEG recordings," *Clinical Neurophysiology*, vol. 119, no. 1, pp. 197–211, 2008.
- [43] L. Lü and T. Zhou, *Link Prediction*, Higher Education Press, Beijing, China, 2013.
- [44] P. Zwoliński, M. Roszkowski, J. Zygierevicz, S. Haufe, G. Nolte, and P. J. Durka, "Open database of epileptic EEG with MRI and postoperational assessment of foci—a real world verification for the EEG inverse solutions," *Neuroinformatics*, vol. 8, no. 4, pp. 285–229, 2010.
- [45] S. I. Dimitriadis, N. A. Laskaris, V. Tsirka, M. Vourkas, S. Micheloyannis, and S. Fotopoulos, "Tracking brain dynamics via time-dependent network analysis," *Journal of Neuroscience Methods*, vol. 193, pp. 145–155, 2010.
- [46] T. Zhou, L. Lu, and Y. C. Zhang, "Predicting missing links via local information," *European Physical Journal B*, vol. 71, no. 4, pp. 623–630, 2009.
- [47] H. E. Yan, Z. Xiao-Ting, and Y. Yun-Li, "Node similarity algorithm on complex network and its application in epilepsy auxiliary diagnosis," *Computer Systems and Applications*, vol. 26, pp. 9–15, 2017.
- [48] M. E. J. Newman, "Assortative mixing in networks," *Physical Review Letters*, vol. 89, no. 20, p. 208701, 2002.
- [49] S. Feldt, P. Bonifazi, and R. Cossart, "Dissecting functional connectivity of neuronal microcircuits: experimental and theoretical insights," *Trends in Neurosciences*, vol. 34, no. 5, pp. 225–236, 2011.
- [50] J. D. Medaglia, M. E. Lynall, and D. S. Bassett, "Cognitive network neuroscience," *Journal of Cognitive Neuroscience*, vol. 27, no. 8, pp. 1471–1491, 2015.
- [51] C. C. Hilgetag and A. Goulas, "Is the brain really a small-world network?," *Brain Structure and Function*, vol. 221, no. 4, pp. 2361–2366, 2016.
- [52] D. S. Bassett, B. G. Nelson, B. A. Mueller, J. Camchong, and K. O. Lim, "Altered resting state complexity in schizophrenia," *NeuroImage*, vol. 59, no. 3, pp. 2196–2207, 2012.
- [53] Q. Ou, Y.-D. Jin, T. Zhou, B.-H. Wang, and B.-Q. Yin, "Power-law strength-degree correlation from resource-allocation dynamics on weighted networks," *Physical Review E*, vol. 75, no. 2, p. 021102, 2007.

## Research Article

# A Method for Tooth Model Reconstruction Based on Integration of Multimodal Images

Xinwen Zhou,<sup>1,2</sup> Yangzhou Gan,<sup>2,3</sup> Jing Xiong,<sup>2</sup> Dongxia Zhang,<sup>2</sup> Qunfei Zhao,<sup>1</sup>  
and Zeyang Xia <sup>2,3</sup>

<sup>1</sup>Department of Automation, Shanghai Jiao Tong University, Shanghai 200240, China

<sup>2</sup>Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China

<sup>3</sup>CAS Key Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institutes of Advanced Technology, Shenzhen 518055, China

Correspondence should be addressed to Zeyang Xia; [zy.xia@siat.ac.cn](mailto:zy.xia@siat.ac.cn)

Received 17 November 2017; Accepted 7 May 2018; Published 20 June 2018

Academic Editor: Yong Xia

Copyright © 2018 Xinwen 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.

A complete digital tooth model is needed for computer-aided orthodontic treatment. However, current methods mainly use computed tomography (CT) images to reconstruct the tooth model which may require multiple CT scans during orthodontic progress, and the reconstructed model is also inaccurate in crown area. This study developed a tooth model reconstruction method based on integration of CT images and laser scan images to overcome these disadvantages. In the method, crown models and complete tooth models are first reconstructed, respectively, from laser scan images and CT images. Then, crown models from laser scan images and tooth models from CT images are registered. Finally, the crown from laser scan images and root from CT images were fused to obtain a new tooth model. Experimental results verified that the developed method is effective to generate the complete tooth model by integrating CT images and laser scan images. Using the proposed method, the reconstructed models provide more accurate crown than CT images, and it is feasible to obtain complete tooth models at any stage of orthodontic treatment by using one CT scan at the pretreatment stage and one laser scan at that stage to avoid multiple CT scans.

## 1. Introduction

In clinical orthodontics, a digital three-dimensional (3D) complete tooth model is needed for diagnosis, treatment planning, appliance design, tooth movement monitoring, and so on. With the development of the imaging and computer techniques, it is feasible to reconstruct digital tooth models from 3D dental images to conduct computer-aided orthodontic treatment.

Currently, laser scan images and computed tomography (CT) images are the most widely used 3D images in clinical orthodontics. Laser scan images have a high resolution up to ten-micron level, and the reconstructed tooth models have been used in space analysis, diagnosis, and computer-aided design of personalized orthodontic appliance, and so on [1–3]. However, the laser scan images only provide the 3D information of crown surface, and CT images is necessary in these applications where 3D information of root is needed

including tooth arrangement, tooth movement monitoring, and orthodontic treatment simulation [4–7]. Complete 3D tooth models can be reconstructed from CT images [8, 9]. However, the reconstructed model from CT images is inaccurate in the crown area to be employed for personalized orthodontic appliance design due to the low image resolution [10]. In addition, multiple CT scans are needed in clinic for progress and posttreatment records, which is not recommended, since the subject would be exposed to high level of radiation [10, 11].

In this study, a new tooth model reconstruction method based on integration of laser scan images and CT images was developed. Crown models and complete tooth models were first reconstructed, respectively, from laser scan images and CT images. The crown models and tooth models were then registered to align the crown part of the two types of models. New complete tooth models were finally generated by fusing the crown from laser scan images and root from CT images.

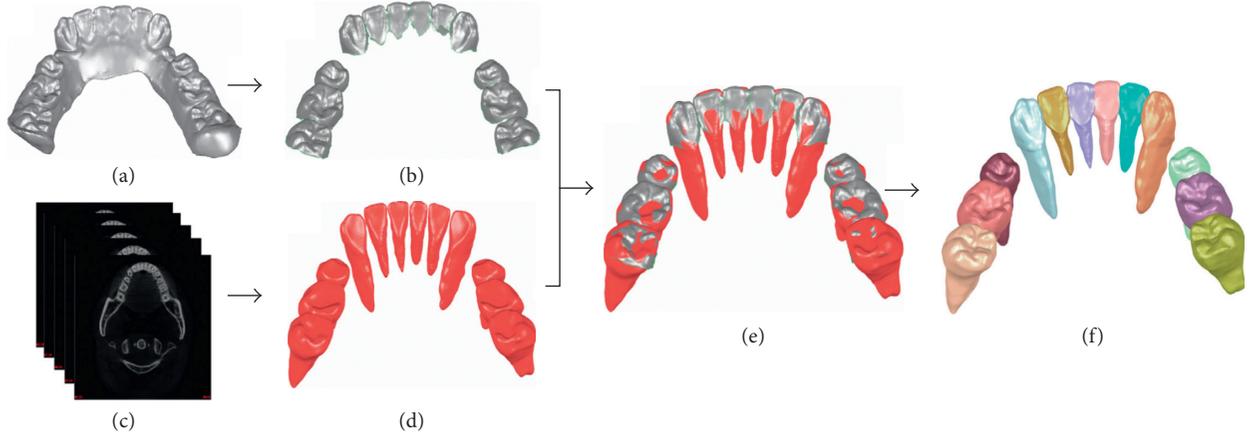


FIGURE 1: Framework of the tooth model reconstruction method based on integration of laser scan images and CT images. (a) Laser scan images of mandibular; (b) individual crowns segmented from laser scan images; (c) CT images of mandibular; (d) 3D tooth models reconstructed from CT images; (e) registration of crown models from laser scan images and CT images; (f) tooth models reconstruction by fusing crowns from the laser scan images and roots from CT images.

The fusion of laser scan images and CT images is challenging. It is difficult to extract the virtual root boundary (the boundary to be combined with the crown from laser scan images) from the CT tooth model because of the complex boundary of laser crown. Additionally, it is difficult to combine the registered crown and root models since the density of vertexes of the two models is inconsistent due to different image spatial resolutions of two sources.

Compared to the previous tooth model reconstruction method from laser scan images or CT images, the contribution of this study mainly includes two aspects. First, the reconstructed tooth model using the developed method can provide the complete 3D information of tooth, and the crown part is accurate enough to be applied for appliance design. Second, by applying the developed method, it is feasible to obtain the complete tooth model at any stage of orthodontic treatment by using one CT scan at pretreatment stage and one laser scan at that stage to avoid multiple CT scans.

## 2. Methods

The framework of the tooth model reconstruction method based on integration of laser scan images and CT images is shown in Figure 1. The laser crown models (Figure 1(b)) and completed CT tooth models (Figure 1(d)) are firstly segmented and reconstructed from laser scan images (Figure 1(a)) and CT images (Figure 1(c)), respectively. The two types of models are then registered (Figure 1(e)) using principal component analysis (PCA) algorithm [12] and iterative closest point (ICP) algorithm [13] to align their crown parts. Finally, the crowns from laser scan images and roots from CT images are fused to generate new tooth models (Figure 1(f)) using the Delaunay-based region-growing method [14].

**2.1. Crown Model Segmentation from Laser Scan Images.** The triangular mesh model (Figure 2(a)) including all tooth crown surface in stereolithographic format is automatically obtained from the laser scanner. Individual tooth crown models

(Figure 2(d)) are segmented from the mesh model by applying a modified fast watershed mesh segmentation method.

In the commonly used fast watershed mesh segmentation method [15], a height function is defined based on the curvature of triangular facet to extract the boundary of neighboring mesh models. However, over segmentation may occur when using this method, and the method may fail to segment individual crowns due to the slow curvature change at the boundary between neighboring crowns. In this study, a modified fast watershed mesh segmentation algorithm is developed to segment individual crowns. Compared to the commonly used fast watershed mesh segmentation method, the modification mainly includes the following two points. (1) A region-growing algorithm [16] is applied to presegment the occlusion area of each crown to avoid over segmentation of the commonly used fast watershed mesh segmentation method (Figure 2(b)). (2) Both curvature and area of triangular facet are employed to define a height function to extract boundary of neighboring crowns (Figure 2(c)). For a given triangular facet  $t_1$ , the height function  $H(t_1, t_2)$  between  $t_1$  and one of its first-order neighboring triangular facets  $t_2$  is written as

$$H(t_1, t_2) = w \frac{\text{area}(t_1) + \text{area}(t_2)}{\max(\text{area}(t_i))} + (1 - w)C(t_1, t_2), \quad (1)$$

where  $t_i$  represents the 1-neighboring facets of  $t_1$ ,  $\text{area}(\cdot)$  is the area of a facet,  $C(t_1, t_2)$  is the curvature of  $t_1$  and  $t_2$  [15], and  $w$  is the weight factor which could be adjusted according to the model. In this study,  $w$  is empirically set to be 0.16.

### 2.2. Tooth Segmentation and Reconstruction from CT Images.

To reconstruct the complete tooth model from CT images (Figure 3(a)), tooth contours (Figure 3(c)) are first segmented from transverse section slice-by-slice (Figure 3(b)) using the hybrid level set-based method [8, 9], and 3D tooth surface model (Figure 3(d)) is reconstructed from the segmented tooth contours using the Marching Cube algorithm [17]. In the procedure of tooth contour segmentation

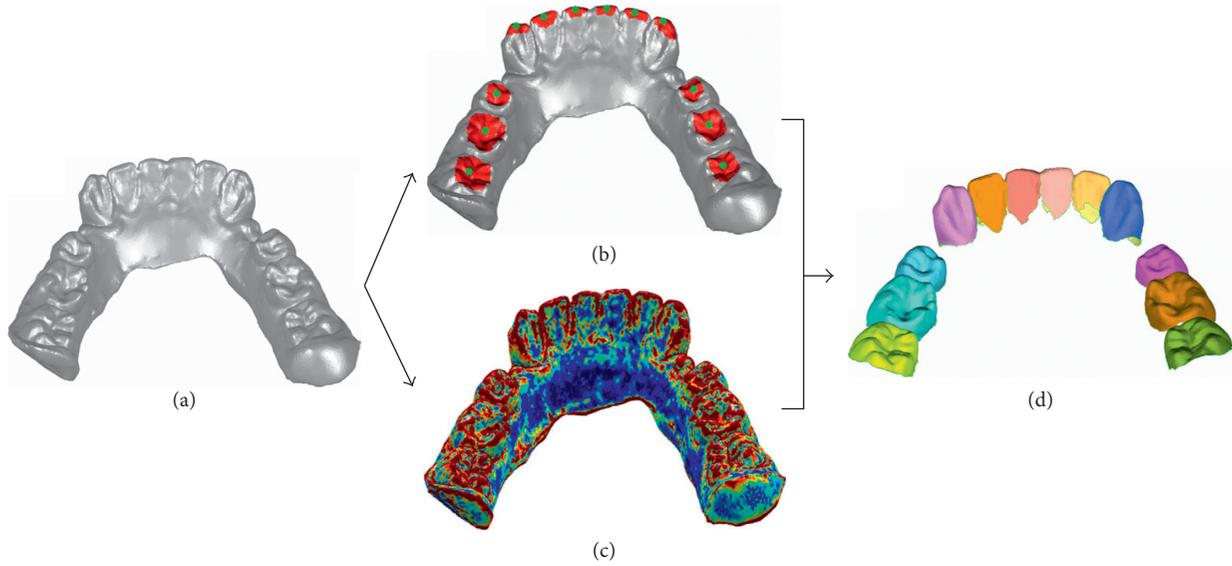


FIGURE 2: Modified fast watershed mesh segmentation method for individual crown model segmentation. (a) Mesh model from laser scanning of mandibular; (b) manually selected seed points of each crown (green points) and presegmentation result in the occlusion area of each crown; (c) height function distribution of facets; (d) individual crown segmented using the modified fast watershed mesh segmentation method.

(Figure 3(b)), a user first manually selects a starting slice from the crown part of the volumetric CT images and picks seed points for each tooth in this slice. Then, the tooth contour is segmented slice-by-slice automatically from the CT images. The automatic segmentation starts from the selected starting slice and propagates along the crown and root directions for crown and root segmentation, respectively. For each slice, the Radon transform is employed to extract a separation line of neighboring teeth [9], and the hybrid level set model is then applied to segment each tooth contour from the mesial and distal sides of the corresponding separation line. Tooth contour propagation strategy which uses segmented tooth contour of previous slice as the tooth shape prior to current slice is employed to initialize the tooth contour automatically. More details of the hybrid level set-based tooth contour segmentation method can be found in [8].

**2.3. Registration of Crown Models from CT Images and Laser Scan Images.** In this study, the aim of model registration is to align the crown part of the two types of models. For the simplicity of computation, the tooth model reconstructed from CT images is sectioned using a plane to generate a crown model for registration with the crown model from laser scan images. The registration between the two types of models is performed through two steps: a coarse registration step based on PCA algorithm [12] and a fine registration step based on ICP algorithm [13]. During the registration process, the crown models from laser scan images are fixed, and the crown models from CT images are registered to the laser scan crown models.

In the coarse registration using PCA algorithm, the covariance matrix of each model is calculated, respectively, from the corresponding node coordinates. Three orthogonal unit eigenvectors are then extracted from the covariance

matrixes, respectively, to establish the body-fitted coordinate system of the corresponding mesh model, and the origin of the coordinate system is set at the centroid of all the mesh model nodes. The aim of coarse registration is to find an affine transformation matrix such that the two body-fitted coordinate systems can be aligned after the affine transformation. Let  $CM_{laser}$  and  $CM_{CT}$  be the covariance matrixes of crown models from laser scan images and CT images, respectively, and  $EM_{laser}$  and  $EM_{CT}$  be the corresponding eigenvector matrixes. The rotation matrix  $R_1$  and the translation matrix  $T_1$  of the affine transformation can be obtained as follows:

$$\begin{aligned} R_1 &= EM_{laser} \cdot \text{Inv}(EM_{CT}), \\ T_1 &= \text{Centroid}(P_{laser}) - R_1 \cdot \text{Centroid}(P_{CT}), \end{aligned} \quad (2)$$

where  $\text{Inv}(\cdot)$  is the matrix inverse operator,  $P_{CT}$  and  $P_{laser}$  are the node sets of the two mesh models, respectively, and  $\text{Centroid}(\cdot)$  is the centroid of the node set operator. Then an affine transformation is performed on  $P_{CT}$  using  $R_1$  and  $T_1$ , to generate the coarse registration result  $P_{CCT}$ .

The fine registration using ICP algorithm aims at finding a set of affine transformation matrixes such that the mean square error (MSE) of the distance between the corresponding nodes of the laser scan crown and the coarsely registered CT crown models achieve minimum after the affine transformations. Let  $M$  and  $N$  denote the number of nodes in  $P_{CCT}$  and  $P_{laser}$ , respectively. The procedure of the ICP algorithm for the fine registration is conducted as follows.

Step 1. Build the  $kd$ -tree [18] of  $P_{CCT}$  and  $P_{laser}$ , respectively.

Step 2. Calculate the rotation matrix  $R_{int}$  and translation matrix  $T_{int}$  using the quaternion method [13].

Step 3. Transform  $P_{CCT}$  using an affine transformation with parameters  $R_{int}$  and  $T_{int}$ .

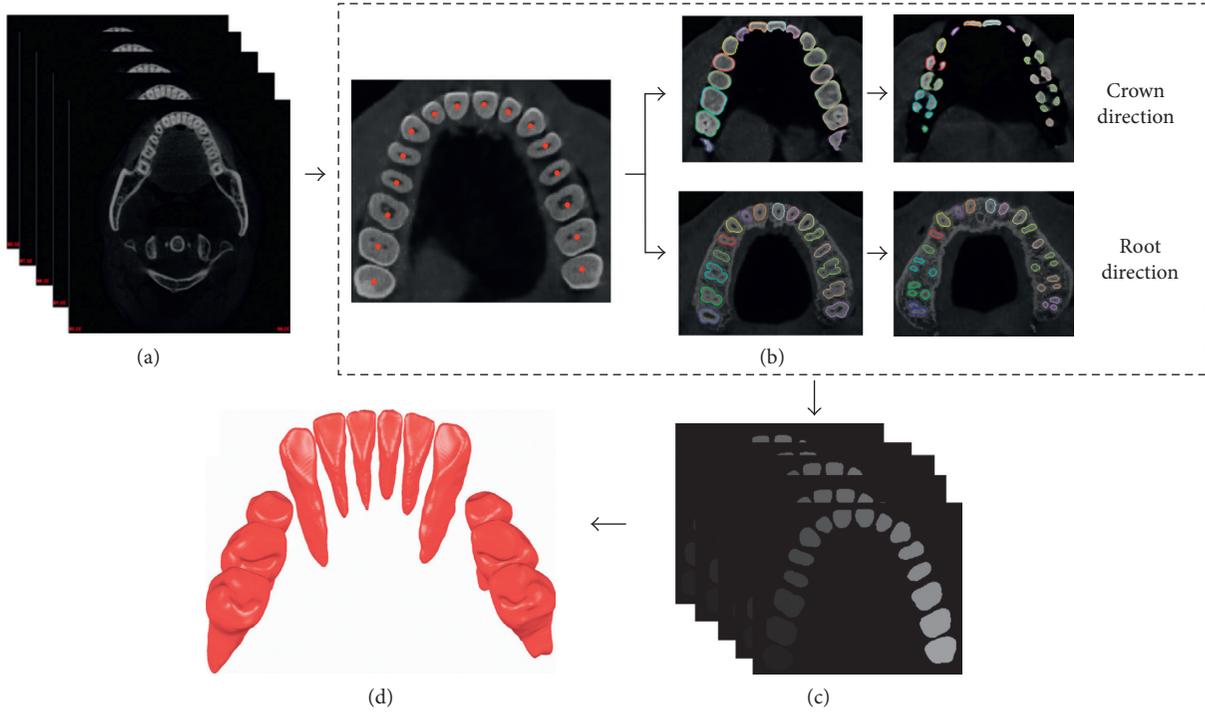


FIGURE 3: Framework of tooth segmentation and reconstruction from CT images. (a) CT images; (b) tooth contour segmentation slice-by-slice; (c) segmented individual tooth region in each slice; (d) reconstructed tooth models from the segmented tooth region using marching cube algorithm.

Step 4. For each node  $n_i$  ( $i = 1, \dots, M$ ) in  $P_{CCT}$ , search its nearest node  $n_j$  ( $j = 1, \dots, N$ ) in  $P_{laser}$  and calculate the corresponding Euclidean distance  $D_{ij}$  between  $n_i$  and  $n_j$ . The registration error  $E_C$  is defined as

$$E_C = \sqrt{\frac{\sum_{i=1}^M D_{ij}^2}{M}}, \quad (3)$$

Step 5. Go back to step 2 until the registration error  $E_C$  is smaller than a preset threshold  $E_{th}$ .

**2.4. Fusion of Crown from Laser Scan Images and Root from CT Images.** In this study, the Delaunay-based region-growing (DBRG) algorithm [14] is applied to generate a new tooth model by fusing the crown models from laser scan images and root models from CT images. In this procedure, each tooth model is proceeded independently using the DBRG algorithm. Two sets of triangle denoted by  $F$  and  $R_Q$ , respectively, and a set of edges denoted by  $E$  are defined. The fusion of the laser scan crown and CT root models is performed as follows:

Step 1. Calculate the Delaunay triangulation  $T$  of all the vertexes in the two registered models.

Step 2. Choose a starting triangle from  $T$ , put the starting triangle into  $F$ , and put its edges into  $E$  (the starting triangle is selected from those triangles with a largest  $z$  coordinate of vertexes and minimum circumradius).

Step 3. Calculate the local smooth degree (LSD) of these triangles in  $T$  who have edges in  $E$  and put these triangles into  $R_Q$  (the definition of LSD can be found in [14]).

Step 4. Denote the triangle with the largest LSD in  $R_Q$  by  $t_0$  and check whether the local geometry and topology between  $t_0$  and triangles in  $F$  is correct. If it is correct, (a) remove  $t_0$  from  $R_Q$ , (b) put  $t_0$  into  $F$ , (c) put edges of  $t_0$  into  $E$ , and (d) delete the edges in  $E$  that are no longer the boundary edges of  $F$ . If not, remove  $t_0$  from  $R_Q$  and repeat Step 4 until  $R_Q$  is empty.

Step 5. Go back to Step 3 until  $R_Q$  is empty.

### 3. Experiments

This study was reviewed and approved by Institutional Review Board of Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. Written informed consents of the subjects are obtained. Intraoral laser scan images and CBCT images of five subjects (4 male, 1 female; average age  $18 \pm 1.9$ ) with crowded teeth in need of orthodontic treatment are obtained at the pretreatment stage. The laser scan images and CBCT images have isotropic voxel sizes of  $50 \mu\text{m}$  and  $0.125 \text{ mm}$ , respectively.

**3.1. Results of Model Reconstruction from 3D Images.** Tooth models of one subject reconstructed from laser scan images and CBCT images are shown in Figures 4(a) and 4(b), respectively. Visually, the crown models reconstructed from laser scan images provide more detail information of crown than those from CBCT images.

**3.2. Results of Crown Model Registration.** The superimposed crown models from laser scan images and CBCT images of

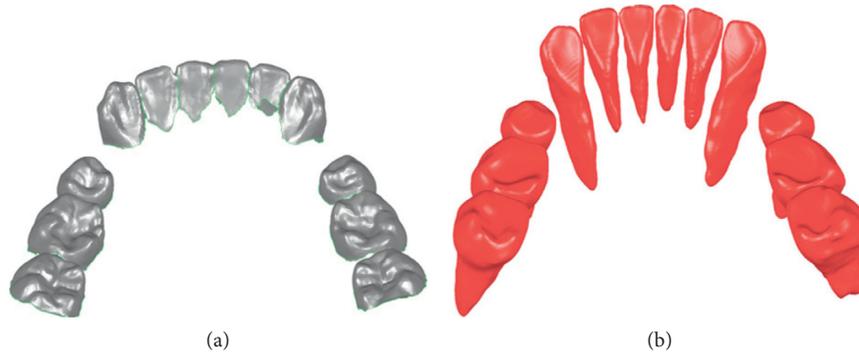


FIGURE 4: Crown and tooth model reconstruction results of one subject. (a) Crown models reconstructed from laser scan images; (b) tooth models reconstructed from CBCT images.

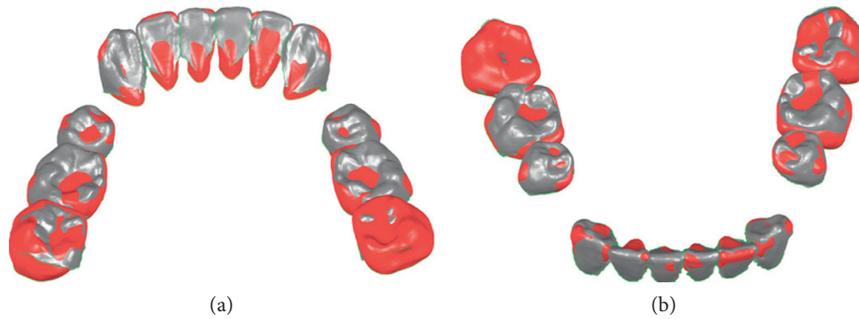


FIGURE 5: Crown registration results from lingual view (a) and labial view (b). Red: the crown model based on CT images; gray: the crown model based on laser scan images.

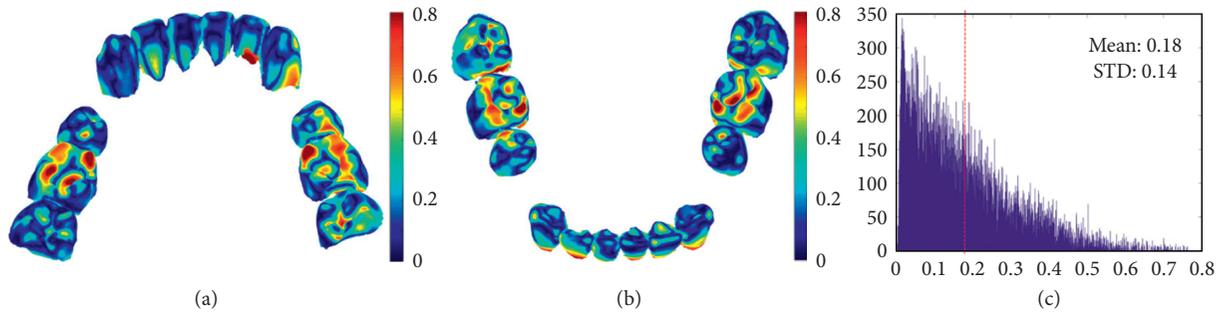


FIGURE 6: Error map of registration result from lingual view (a) and labial view (b), and the histogram of distance between the corresponding nodes of two model (c).

one subject after model registration are shown in Figure 5, and the corresponding registration error distribution of nodes is shown in Figure 6.

In this study, the average distance (AD) from the laser scan crown model to the registered CT crown model is applied to quantify the registration error of the two models. The registration error for all the tested subjects is  $0.19 \pm 0.03$  mm, and the registration error of each subject is listed in Table 1.

TABLE 1: Registration error of each subject.

Subjects	AD (mm)
Subject 1	0.18
Subject 2	0.16
Subject 3	0.23
Subject 4	0.21
Subject 5	0.17
All	$0.19 \pm 0.03$

**3.3. Results of Model Fusion.** For all the tested images, new tooth models are successfully generated based on the fusion of the crown from laser scan images and the root from CT images. Figure 7 shows the generated tooth models of one

subject based on the fusion. Compared to the models directly reconstructed from laser scan images and CT images in Figure 4, the models obtained by the fusion method not only contain the complete 3D tooth information but also provide accurate crown.

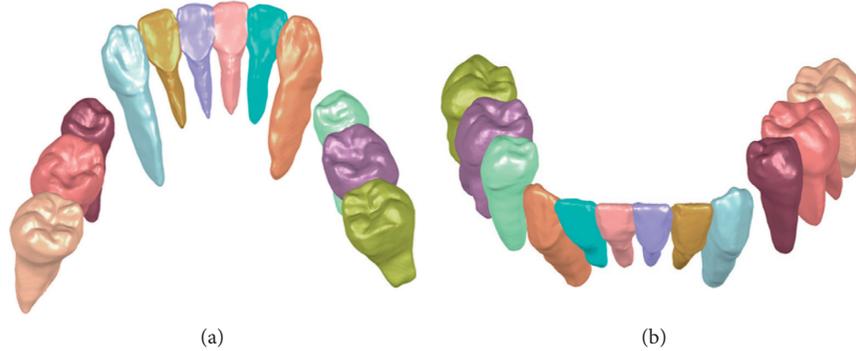


FIGURE 7: Tooth models from the fusion of crown from laser scan images and root from CT images. (a) Lingual view of fused tooth models; (b) labial view of fused tooth models.

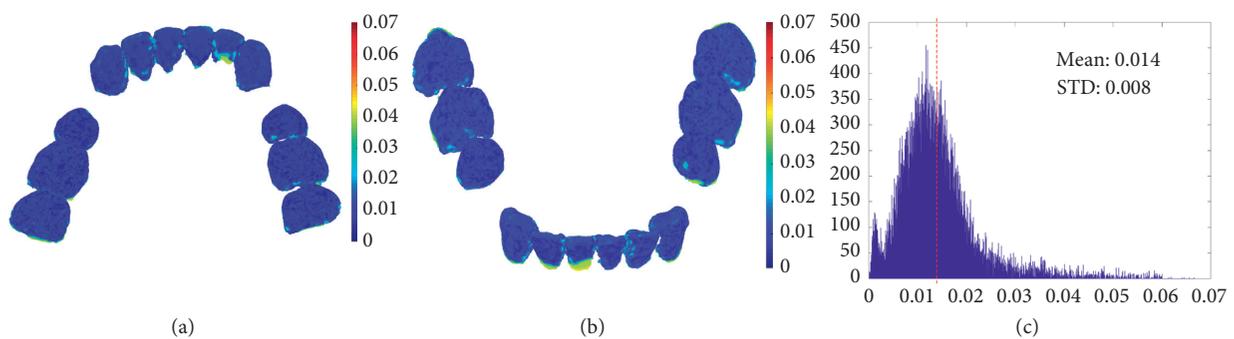


FIGURE 8: Error between the laser scan crown model and the tooth models from the fusion procedure. (a) Lingual view; (b) labial view; (c) histogram of the error.

#### 4. Discussion

In computer-aided orthodontics, complete 3D digital tooth models are needed for diagnosis, treatment planning, and treatment simulation. One can reconstruct the complete tooth models from CT images. However, the reconstructed models are inaccurate in the crown part for personalized appliance design. Additionally, multiple CT scans are necessary for progress and posttreatment records if only using the CT image to reconstruct the tooth models. Compared with CT images, laser scan images can be used to reconstruct much more accurate crown models without radiation [19, 20]. In this study, we developed a new tooth model reconstruction method based on the fusion of the crown model from laser scan images and root model from CT images. By applying the developed method, it was feasible to obtain the complete tooth model at any stage of orthodontic treatment by using one CT scan at pretreatment stage and one laser scan at that stage to avoid multiple CT scans. Additionally, the reconstructed tooth model using the developed method provided more accurate crown than CT images.

In the developed method, the crown models from laser scan images and the tooth models from CT images are first registered and then fused to generate the complete tooth models, and both the registration and fusion procedure would affect the accuracy of the reconstructed model. In previous works of the dental model registration, the mean registration errors are among 0.1 mm and 0.3 mm [21–23]. In this study, the mean registration error in the developed

TABLE 2: Fusion error of each subject.

Subjects	AD (mm)
Subject 1	0.014
Subject 2	0.011
Subject 3	0.020
Subject 4	0.017
Subject 5	0.013
All	$0.015 \pm 0.004$

method is 0.19 mm and is clinically acceptable [24]. To evaluate the model fusion error, the laser scan crown is chosen as the ground truth and compared with the fused tooth model, and AD was used to quantify the fusion error. The error map of one subject is shown in Figure 8. For all the tested subjects, the fusion error is  $0.015 \pm 0.004$  mm which is rather small and clinically acceptable [25], and the fusion error of each subject is listed in Table 2.

In the developed method, manual initialization is needed for the segmentation of CT images and laser scan images which led to limited difference of segmentation results for different trained users [8, 15]. While the registration and fusion procedure are performed fully automatically, we could thus achieve results with good reproducibility and reliability compared to those manually operating works [21, 25]. The proposed method may fail to reconstruct angled teeth since the slice-by-slice image segmentation strategy used in the method has difficulty in the segmentation of these teeth [26].

## 5. Conclusion

This study presented a new tooth model reconstruction method based on integration of laser scan images and CBCT images. Compared to those commonly used tooth model reconstructed methods which directly reconstruct complete tooth model from CT images, the proposed method can generate tooth model with more accurate crown and can obtain a complete tooth model at any stage of orthodontic treatment by using one CT scan at pretreatment stage and one laser scan at that stage to avoid multiple CT scans. Thus, the proposed tooth model reconstruction method based on image integration can benefit the computer-aided orthodontic treatment.

## Conflicts of Interest

The authors declare no conflicts of interest.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China (61601452), Guangdong Natural Science Funds for Distinguished Young Scholar (2015A030306020), Youth Innovation Promotion Association, Chinese Academy of Sciences (2015301), and Shenzhen Research Project (JCYJ20170818162801483, JCYJ20170413162458312, GJHS20160331185913023, and GJHS20170314154158554).

## References

- [1] S. Mullen, C. Martin, P. Ngan, and M. Gladwin, "Accuracy of space analysis with emodels and plaster models," *American Journal of Orthodontics and Dentofacial Orthopedics*, vol. 132, no. 3, pp. 346–352, 2007.
- [2] N. Motohashi and T. Kuroda, "A 3D computer-aided design system applied to diagnosis and treatment planning in orthodontics and orthognathic surgery," *European Journal of Orthodontics*, vol. 21, no. 3, pp. 263–274, 1999.
- [3] L. Liu, B. He, J. Zhuang, L. Zhang, and A. Lv, "Force measurement system for invisalign based on thin film single force sensor," *Measurement*, vol. 97, pp. 1–7, 2017.
- [4] H. Tong, D. Kwon, J. Shi, N. Sakai, R. Enciso, and G. Sameshima, "Mesiodistal angulation and faciolingual inclination of each whole tooth in 3-dimensional space in patients with near-normal occlusion," *American Journal of Orthodontics and Dentofacial Orthopedics*, vol. 141, no. 5, pp. 604–617, 2012.
- [5] A. Yagci, I. Veli, T. Uysal, F. Ucar, T. Ozer, and S. Enhos, "Dehiscence and fenestration in skeletal class I, II, and III malocclusions assessed with cone-beam computed tomography," *Angle Orthodontist*, vol. 82, no. 1, pp. 67–74, 2012.
- [6] Z. Librizzi, Z. Kalajzic, D. Camacho, S. Yadav, R. Nanda, and F. Uribe, "Comparison of the effects of three surgical techniques on the rate of orthodontic tooth movement in a rat model," *Angle Orthodontist*, vol. 87, no. 5, pp. 717–724, 2017.
- [7] J. Gao, W. Xu, and Z. Ding, "3D finite element mesh generation of complicated tooth model based on CT slices," *Computer Methods and Programs in Biomedicine*, vol. 82, no. 2, pp. 97–105, 2006.
- [8] Y. Gan, Z. Xia, J. Xiong, Q. Zhao, Y. Hu, and J. Zhang, "Toward accurate tooth segmentation from computed tomography images using a hybrid level set model," *Medical Physics*, vol. 42, no. 1, pp. 14–27, 2015.
- [9] Z. Xia, Y. Gan, J. Xiong, Q. Zhao, and J. Chen, "Crown segmentation from computed tomography images with metal artifacts," *IEEE Signal Process Letters*, vol. 23, no. 5, pp. 678–682, 2016.
- [10] G. Swennen, E. Barth, C. Eulzer, and F. Schutyser, "The use of a new 3D splint and double CT scan procedure to obtain an accurate anatomic virtual augmented model of the skull," *International Journal of Oral and Maxillofacial Surgery*, vol. 36, no. 2, pp. 146–152, 2007.
- [11] A. Alqerban, G. Willems, C. Bernaerts, J. Vangastel, C. Politis, and R. Jacobs, "Orthodontic treatment planning for impacted maxillary canines using conventional records versus 3D CBCT," *European Journal of Orthodontics*, vol. 36, no. 6, pp. 698–707, 2014.
- [12] I. Jolliffe, *Principal Component Analysis*, Springer, New York, NY, USA, 1986.
- [13] P. Besl and N. McKay, "A method for registration of 3-D shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, 1992.
- [14] C. Kuo and H. Yau, "Delaunay-based region-growing approach to surface reconstruction from unorganized points," *Computer-Aided Design*, vol. 37, no. 8, pp. 825–835, 2005.
- [15] C. Li, C. Zhang, and G. Wang, "Fast marker-controlled interactive mesh segmentation," *Acta Scientiarum Naturalium Universitatis Pekinensis*, vol. 42, no. 5, pp. 662–667, 2006, in Chinese.
- [16] R. Adams and L. Bischof, "Seeded region growing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 6, pp. 641–647, 1994.
- [17] W. E. Lorensen and H. Cline, "Marching cubes: a high resolution 3D surface construction algorithm," *ACM SIGGRAPH Computer Graphics*, vol. 21, no. 4, pp. 163–169, 1987.
- [18] I. Wald and V. Havran, "On building fast kd-trees for ray tracing, and on doing that in  $O(N \log N)$ ," in *Proceedings of the IEEE Symposium on Interactive Ray Tracing*, pp. 61–69, Lake Tahoe, NV, USA, November 2006.
- [19] D. R. Stevens, C. Flores-Mir, B. Nebbe, D. W. Raboud, G. Heo, and P. W. Major, "Validity, reliability, and reproducibility of plaster vs. digital study models: comparison of peer assessment rating and Bolton analysis and their constituent measurements," *American Journal of Orthodontics and Dentofacial Orthopedics*, vol. 129, no. 6, pp. 794–803, 2006.
- [20] C. Kau, S. Richmond, A. Zhurov et al., "Reliability of measuring facial morphology with a 3-dimensional laser scanning system," *American Journal of Orthodontics and Dentofacial Orthopedics*, vol. 128, no. 4, pp. 424–430, 2005.
- [21] H.-H. Lin, W.-C. Chiang, L.-J. Lo, S. S.-P. Hsu, C.-H. Wang, and S.-Y. Wan, "Artifact-resistant superimposition of digital dental models and cone-beam computed tomography images," *Journal of Oral and Maxillofacial Surgery*, vol. 71, no. 11, pp. 1933–1947, 2013.
- [22] J. Uechi, M. Okayama, T. Shibata et al., "A novel method for the 3-dimensional simulation of orthognathic surgery by using a multimodal image-fusion technique," *American Journal of Orthodontics and Dentofacial Orthopedics*, vol. 130, no. 6, pp. 786–798, 2006.
- [23] G. Swennen, M. Mommaerts, J. Abeloos et al., "A cone-beam CT based technique to augment the 3D virtual skull model with a detailed dental surface," *International Journal of Oral and Maxillofacial Surgery*, vol. 38, no. 1, pp. 48–57, 2009.
- [24] C. Dietrich, A. Ender, S. Baumgartner, and A. Mehl, "A validation study of reconstructed rapid prototyping models

- produced by two technologies,” *Angle Orthodontist*, vol. 87, no. 5, pp. 782–787, 2017.
- [25] T. Kihara, K. Tanimoto, M. Michida et al., “Construction of orthodontic setup models on a computer,” *American Journal of Orthodontics and Dentofacial Orthopedics*, vol. 141, no. 6, pp. 806–813, 2012.
- [26] Y. Gan, Z. Xia, J. Xiong, G. Li, and Q. Zhao, “Tooth and alveolar bone segmentation from dental computed tomography images,” *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 1, pp. 196–204, 2018.

## Research Article

# Bayesian Classification Models for Premature Ventricular Contraction Detection on ECG Traces

**Manuel M. Casas,<sup>1</sup> Roberto L. Avitia<sup>1</sup> ,<sup>1</sup> Felix F. Gonzalez-Navarro,<sup>2</sup> Jose A. Cardenas-Haro,<sup>3</sup> and Marco A. Reyna<sup>2</sup>**

<sup>1</sup>Facultad de Ingenieria, Universidad Autonoma de Baja California, Mexicali, BC, Mexico

<sup>2</sup>Instituto de Ingenieria, Universidad Autonoma de Baja California, Mexicali, BC, Mexico

<sup>3</sup>Computer and Electrical Engineering and Computer Science Department, California State University, Bakersfield, CA, USA

Correspondence should be addressed to Roberto L. Avitia; [ravitia@uabc.edu.mx](mailto:ravitia@uabc.edu.mx)

Received 23 August 2017; Revised 17 December 2017; Accepted 1 April 2018; Published 10 May 2018

Academic Editor: Yong Xia

Copyright © 2018 Manuel M. Casas 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.

According to the American Heart Association, in its latest commission about Ventricular Arrhythmias and Sudden Death 2006, the epidemiology of the ventricular arrhythmias ranges from a series of risk descriptors and clinical markers that go from ventricular premature complexes and nonsustained ventricular tachycardia to sudden cardiac death due to ventricular tachycardia in patients with or without clinical history. The premature ventricular complexes (PVCs) are known to be associated with malignant ventricular arrhythmias and sudden cardiac death (SCD) cases. Detecting this kind of arrhythmia has been crucial in clinical applications. The electrocardiogram (ECG) is a clinical test used to measure the heart electrical activity for inferences and diagnosis. Analyzing large ECG traces from several thousands of beats has brought the necessity to develop mathematical models that can automatically make assumptions about the heart condition. In this work, 80 different features from 108,653 ECG classified beats of the gold-standard MIT-BIH database were extracted in order to classify the Normal, PVC, and other kind of ECG beats. Three well-known Bayesian classification algorithms were trained and tested using these extracted features. Experimental results show that the F1 scores for each class were above 0.95, giving almost the perfect value for the PVC class. This gave us a promising path in the development of automated mechanisms for the detection of PVC complexes.

## 1. Introduction

According to the World Health Organization, cardiovascular diseases (CVD) are the main cause of death worldwide. An estimated 17.5 million people died from CVD in 2012, representing 31 of all global deaths [1]. The latest standard in the American Heart Association (AHA) on ventricular arrhythmias and sudden cardiac death in 2006, the epidemiology of ventricular arrhythmias includes a series of risk factors and clinical applications. These arrhythmias range from premature complexes, ventricular tachycardia and sustained ventricular tachycardia in individuals without cardiac issues background to sudden death due to ventricular tachyarrhythmia [2]. The electrocardiogram (ECG) is the main tool for the prediagnosis of heart diseases. Today, computer-aided analysis of short time

ECG records, taken from supine positions, is a well-established procedure.

A normal heartbeat (NB) reflects a heart regular activity condition. On the other hand, premature ventricular contraction (PVC) is a kind of arrhythmia caused by an ectopic cardiac pacemaker located in the ventricle. PVC is a type of ECG arrhythmias that is identified for presenting anomalies in the normal cardiac rhythm, generating alterations in the heart rate that disrupts the mechanic and electric heart activity due to these delayed contractions (premature). On the ECG, these PVCs are characterized by premature and bizarrely shaped QRS complexes, usually wider than 120 ms, and a T wave larger than usual. A PVC event can be seen in healthy people and/or persons with some cardiac disorders, normally asymptomatic. The bizarrely shaped QRS complexes

can increase the risk of a cardiac arrest and eventually may lead to a sudden cardiac death [3]. A bigger problem, however, is to track the presence and number of arrhythmias over days, weeks, and months. Since cardiologists cannot spend a lot of time in the analysis of millions of heartbeats from an individual, it is necessary to use automated mathematical algorithms to detect these abnormal events [4].

There have been so many improvements in ECG conditioning; some of these are signal-to-noise ratio enhancement, wave detection characteristics, heart rate variability analysis, and ECG patterns classification, among others. Since the new algorithms are increasingly more powerful and precise, gaps between the use of recent algorithms and the standard analysis methodology of the available evidence have begun to emerge [5–7]. Nazarahari et al. [8] used a neural network (NN) to classify six relevant types of heartbeats from a set of features created by new wavelet functions along with different distances and principal component analysis (PCA) to reduce the dimensionality. Martis et al. [9] applied the bispectrum computation in each beat and PCA to create the features that ultimately fed NN and support vector machines (SVMs) algorithms to classify between five types of heartbeats including the normal heartbeats (NBs) and premature ventricular contraction (PVC) beats. Afkhami et al. [10] derived morphological, statistical, and temporal features from the heartbeats amid probability density function extracted from the Gaussian mixture modeling (GMM) parameters to train an ensemble of decision trees. Javadi et al. [11] extracted features using the wavelet transform from key morphological shapes of the ECG and combined negative correlation learning with mixture of experts to train a negatively correlated NNs (neural networks). Kamath [12] used the Teager energy operator to derived nonlinear components in time and frequency; consequently, he fed a NN classifier to make predictions for five different arrhythmia beats. Martis et al. [13] segmented the QRS wave from each beat, derived features using the discrete wavelet transform (DWT), and compared PCA, linear discriminant analysis (LDA), and independent component analysis (ICA), three different dimensionality reduction techniques, to obtain the best method with greater performance classifying heartbeats. Sharma and Ray [14] put every heartbeat through the Hilbert-Huang transform for feature extraction along with other set of features as statistical features, Kolmogorov complexity and weighted mean frequency which served as training for a SVM classifier. Banerjee and Mitra [15] proposed heuristic classification based on the cross wavelet transform of ECG signals to classify between abnormal and normal heartbeats. Oliveira et al. [16] designed a dynamic Bayesian network, and with a threshold set by an expert, it is able to classify between PVC and other kinds of beats.

These works give valuable information about the insights of the ECG nature and classification boundaries of the heartbeats and have a high classification performance. In this work, we compared three generative classifiers to distinguish between NB, PVC, and others. We attempted to simplify the feature extraction and use much more simple Bayesian generative model algorithms, Gaussian Naïve Bayes (GNB),

Gaussian linear discriminant analysis (LDA), and quadratic discriminant analysis (QDA). In the case of the GNB, it assumes independence between features in such a way that every feature is parametrized by univariate  $\mu$  and  $\sigma$ , and the LDA/QDA takes into account the joint distribution of the features, and these are parametrized by  $\mu$  and  $\Sigma$ . These parameters are much more simple to understand, interpret, and correlate with the labels in question and are preferred over a complex hyperparameters as those can give much more detailed information about the characteristics, attributes, or components from the recollected and extracted data.

## 2. Materials and Methods

In this work, we considered the MIT-BIH arrhythmia database, available in the PhysioNet web page [17, 18]. It consists of 48 half-hour signal records of two channel ambulatory ECG recordings, digitized at 360 samples per second with 11-bit resolution over a 10 mV maximum range. The most important part of this database is that it has reference notations at each beat done by expert cardiologists.

All the experiments were performed in MATLAB® 2014. For research purposes, along with the MIT-BIH arrhythmia database, the PhysioNet web page provides a file for every ECG record with a beat classification. We relabeled this database for this work in a similar way as in [8–16]. We kept the original labels of NB and PVC, and we mapped the rest of the heartbeats as “Others Beats” (OB). A total of 74,924 beats are classified as NB, 7129 beats as PVC, and 26,600 as OB. Giving a total of 108,653 beats for multiclass classification purposes. Samples of these beats can be seen in Figure 1.

In Figure 2 is shown the workflow that we followed in this research. First, we extracted every beat from every signal in the database. Then, we put the data through a series of preprocessing steps that include the normalization, the transformed space, and the outlier detection. We proceeded with the experimentation using the generative models and the cross validation for the fair and safe comparison of our results.

*2.1. Beat Extraction.* For feature extraction purposes, we segmented each heartbeat taking a time window of 0.2 ms from R peak backwards and 0.46 from R peak forward, lasting 0.66 ms, which is approximately what a normal beat lasts. In this way, we created a matrix where every row was a heartbeat and the columns represented each sample point.

*2.2. Preprocessing and Extraction of Features from ECG Beats.*

In order to train the classifiers and have a better performance, we processed the signals. The first step was to mean normalize every sample of every signal, and that was done with the following equation:

$$x_{\text{normalized}} = \frac{x_{\text{sample}} - x_{\mu}}{x_{\text{max}} - x_{\text{min}}}, \quad (1)$$

where  $x_{\text{sample}}$  corresponds to the sample in a particular ECG signal,  $x_{\mu}$  is the mean of signal,  $x_{\text{max}}$  is the maximum value of

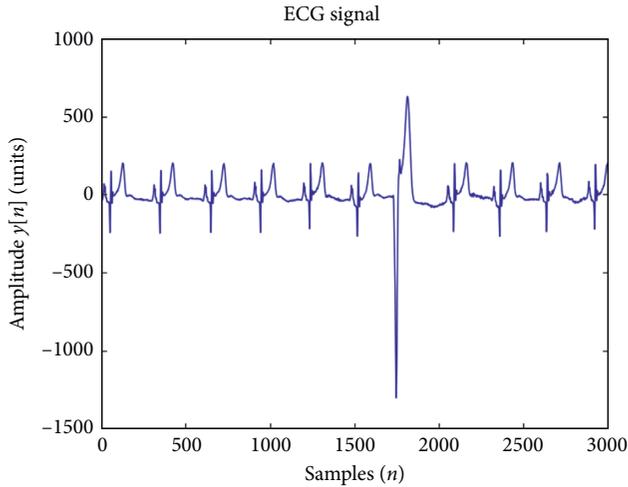


FIGURE 1: Samples of NB and an isolated PVC.

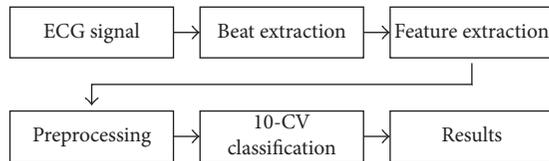


FIGURE 2: Workflow of the experimentation.

ECG signal, and  $x_{\min}$  is the minimum value of the signal where the sample is located.

The main feature extraction is explained in Figure 3. Once the signals were mean normalized, each one was divided in four vectors. Each quartered vector was under a processing procedure for the extraction of 20 features. For each quartered vector, we used a procedure that we named “Feature Statistic Calculation” (FSC) where we calculate the mean, the standard deviation, and the maximum and minimum values. Also, for each quarter, we used another procedure that we named as “Samples Features Extraction” (SFE) where we extracted six samples that characterized the quartered vector. These six characteristic samples are the beginning and the ending of the quartered vector and four samples equally spaced from the quartered vector. The discrete Fourier transform (DFT) was applied in each of these quartered vectors, and for the resulting transformed quartered vectors, the same procedure of implementing the FSC and SFE was performed. At the end, as a result, we have extracted ten features from the quartered vector of the original heartbeat plus another ten features from the transformed quartered vector. As we divided all the signals in four vectors, we ended up with 80 features to feed the classifiers.

A great number of elements that has nothing to do with a disease can distort the ECG signals. The easier ones to remove with digital signal processing techniques (such as filters) are the 60 Hz powerline frequency and muscle-noise signals. However, as these data come from ECG Holters (a wearable device that records the signal from a patient through a considerable number of hours), many of

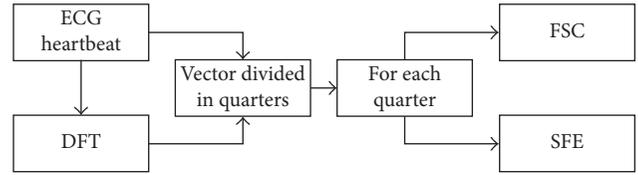


FIGURE 3: Feature extraction.

heartbeats and the signal itself are disturbed by muscle movements, and these heartbeats recorded are very different in shape from the ones recorded before or after by standard methods. Using an outlier algorithm that helped our algorithm to be correctly trained as the models used are based on Gaussian distribution, outliers can affect greatly the performance of the model. The next step in the procedure was to get rid of the samples considered outliers. For this task, an algorithm which is based both in the multivariate Mahalanobis distance and in the comparison of the critical value of the  $\chi^2$  distribution was used [19]. For this, we needed to assume that the dataset behaves as a normal distribution; then, the Mahalanobis distances from every feature sample follow a Chi-square distribution with  $d$  grades of liberty, in this case 80. Any value above the 97.5th quantile is considered an outlier. The dataset was reduced to 84,586 beats, with 58,049 normal beats, 5,222 PVC beats, and 21,315 beats classified as others.

The “cluster-based visualization with scatter matrix” algorithm [20] was implemented in order to visualize and project the data in such way that it could be better classified. The scattering of matrices is a technique used to reduce the dimensionality and to maximize the dispersion between groups. However, this method is for clustered or labeled datasets. It is well known that the total variance  $S_t$  can be scattered in the sum of two terms called scattered matrices, which calculate the variance within the group  $S_w$  and the variances between each group  $S_b$ . Due to the importance of the media within each cluster as representatives of each group, it is natural to project the information in a subspace covered by the media of each group. This can be done defining a set of orthonormal vectors  $\hat{b}_i = 1, \dots, N_c$  ( $N_c$  number of clusters) using the method of Gram-Schmidt orthogonalization:

$$X^c = \sum_{i=1}^{N_c} (X \cdot \hat{b}_i^T) \cdot \hat{b}_i, \quad (2)$$

with  $X$  being the whole dataset. A way to preserve as much separation as possible between the groups, after the projection of the dataset into a subspace covered by the medias of each group, is through the process of whitening before this procedure:

$$X_{\text{esf}} = \alpha \cdot \beta^{-1/2} \cdot X. \quad (3)$$

This last implementation is with the only purpose to reduce the dimensionality and visualization, and it does not intervene or affect in the process of clustering, where alpha and beta are the eigenvectors and eigenvalues, respectively. The computation of the scatter matrices for each class is by the following equations:

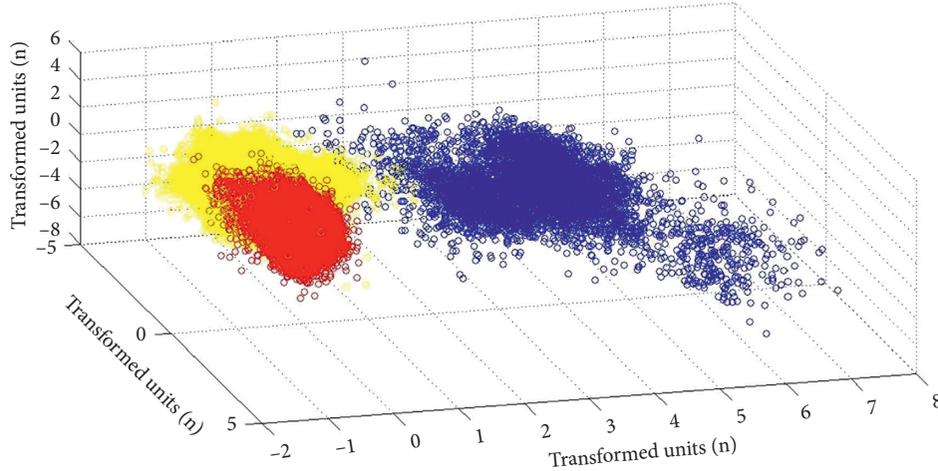


FIGURE 4: Clusters after classifications.

$$\begin{aligned}
 S_w^C &= \sum_{j=1}^{N_c} \sum_{i=1}^{N_j} \left( (X_i^C - \mu_j^C)^T \cdot (X_i^C - \mu_j^C) \right), \\
 S_B^C &= \sum_{j=1}^{N_c} \left( N_j \cdot (\mu_j - \mu)^T \cdot (\mu_j - \mu) \right), \\
 M^C &= (S_w^C)^{-1} \cdot S_B^C,
 \end{aligned} \tag{4}$$

where  $X_i$  is the dataset of  $i$ th cluster,  $\mu$  is the general mean, and  $\mu_i$  is the mean of each cluster. Consequently, the diagonalization of the new scatter matrix  $M^C$  shows that for each class, typically, the major part of the information is within the first eigenvalues. This eigenvector forms the second and third dimension for visualization purposes. The first three lower dimensional projections are shown in Figure 4; this image shows that there are some clear boundaries and three different types of beats, the normal, the PVC, and the others. The dataset was divided uniformly into 70, 15, and 15, for the training set, the validation set, and the testing set, respectively, without repetitions.

**2.3. The Classifiers and the Classification.** Once the data is prepared and ready to be used as a training set for the classifiers, we can make inferences with the Bayesian generative models. The Bayes theorem is defined as

$$P(q_k|X, \theta_k) = \frac{p(X|q_k, \theta_k) \cdot (q_k|\theta_k)}{p(X)}. \tag{5}$$

For classification purposes, the divisor or the marginal probability is not necessary because it plays a constant role, and as we are interested to know which class  $q_k$  have a higher probability, then  $P(q_k|X, \theta_k)$  is proportional to  $p(X|q_k, \theta_k) \cdot (q_k)$  for every class:

$$P(q_k|X, \theta_k) \propto p(X|q_k, \theta_k) \cdot (q_k|\theta_k), \tag{6}$$

where  $P(q_k|X, \theta_k)$  is the posterior probability, in this case, the class  $q_k$  given a sample  $X$  or a feature vector following

a Gaussian distribution with parameters  $\theta_k$ ;  $p(X|q_k, \theta_k)$  is the likelihood of the sample  $X$ , given a class  $q_k$  following a Gaussian distribution with parameters  $\theta_k$ ; and  $p(q_k)$  is the probability that the class  $q_k$  is presented.

For numerical stability, we can use the logarithm operator, and now we can represent the classifier as a sum of logarithms:

$$\log(P(q_k|X, \theta_k)) \propto \log(p(X|q_k, \theta_k)) + \log(p(q_k|\theta_k)). \tag{7}$$

The probability of the priori for each class is given by the ratio  $P = n_c/N$ , where  $n_c$  is number of times  $c$  class is presented, and  $N$  is the total number of instances. This remains constant and only the likelihood changes for every classifier, as it modeled different.

As we have numerical data and in GNB assume that every feature is independent, the likelihood is formed by the multiplication of every feature parameterized by univariate Gaussian distribution's  $\mu$  and  $\sigma$ . LDA and QDA are represented by  $\mu$  and  $\Sigma$ , as these classifiers model the joint distribution of the features, and the likelihood term is modeled by the multivariate Gaussian distribution (MGD) with  $d$  dimensions and  $k$  classes and is defined by the following equation:

$$\text{MGD} = -\frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1} (x-\mu_k)}. \tag{8}$$

And substituting it in likelihood in (7), we have

$$\begin{aligned}
 q_k(x) &= -\frac{1}{2} \log \left| \Sigma_k \right| - \frac{1}{2} (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) \\
 &+ \log(p(q_k|\theta_k)),
 \end{aligned} \tag{9}$$

which leads to the QDA classifier.

If we assume the same covariance for all the classes, this enables us to use a linear classifier for each class:

$$q_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log(p(q_k|\theta_k)). \tag{10}$$

**2.4. Evaluation of the ECG Classifier.** Before explaining the experimentation process, the evaluation needs to be explained. The efficiency of a test is entirely captured by the following four basic measurements: true positive (TP), false negative (FN), false positive (FP), and true negative (TN). From these four basic measurements, all the other statistical measures can be derived. In this context, a true positive means that a PVC was predicted and the arrhythmia really happened, while a true negative means that a PVC was not diagnosed and the arrhythmia was, indeed, not present. However, a false positive means that a PVC was identified, but it really did not happen. Finally, a false negative is that a PVC was not detected although the arrhythmia really was there. Sensitivity (Se) or Recall indicates the ability of a test to identify positive cases; a test with high sensitivity has few false negatives. Positive predictive value (PPV) or Precision provides the probability of being true positive when the test is positive. Equations (11) and (12) show how to calculate the above-mentioned measurements:

$$Se = \frac{TP}{TP + FN} \quad (11)$$

$$PPV = \frac{TP}{TP + FP} \quad (12)$$

For the relation between these two parameters, we used the F1 score implemented as a good evaluator in [21] and can be expressed as

$$F = \frac{2 \cdot Se \cdot PPV}{Se + PPV} \quad (13)$$

In the worst-case scenario, the F1 score is zero if the two parameters are zero; and in the best-case scenario, the F1 score is 1 if the two parameters are one. This ratio gives a good sense of how the algorithm does the classification.

**2.5. Experimentation Set Up.** The experiment was performed as follows: 10-crossfold validation performed for the classifier selection and assessment. The training and validation set were concatenated, and 10-crossfold validation was done for the three classifiers. We calculated the F1 score for every fold, and the mean was extracted in order to have an average of the performance from each classifier. Every class has a F1 score; in this way, every classifier has three F1 scores, and as a way to select the best among these, we calculated once again the mean from those results, and the model with the higher average was selected. With the training and validation set concatenated, the trained classifier chosen was tested then with the test set. The results from this experiment are shown in the following section.

### 3. Results and Discussion

Each record of the MIT-BIH Arrhythmia database was downloaded in the “mat” format. The extraction of every NB, PVC, and OB were just explained in the previous section. A dataset of 108,653 beats with 80 features were extracted, and after taking out the outlier samples, the database was reduced to 84,586 samples. For the training we used 70% of the

TABLE 1: Performance of each classifier tested.

Mean F1 scores for each train-validation test		
GNB	QDA	LDA
0.868	0.983	0.960

TABLE 2: Performance of the QDA classifier.

Performance of the final QDA model			
	Se	PPV	Fscore
NB	0.991	0.987	0.989
OB	0.959	0.974	0.967
PVC	1	0.980	0.990

TABLE 3: Confusion matrix of the classifier selected over the test set.

Confusion matrix of the final QDA model			
	NB	OB	PVC
NB	8698	78	1
OB	114	3014	14
PVC	0	0	769

data, for validation 15% of the data, and the remaining 15% was for testing purposes.

Three supervised learning algorithms were implemented to classify the ECG beats, and the results from this are shown in Table 1. It can be seen that the mean F1 scores for the Naïve Bayes is the lowest with a considerable 0.86. The top ones were the LDA and QDA, with mean F1 scores of 0.96 and 0.98, respectively. Obviously the QDA algorithm has the highest performance.

The new training set is composed of the training and validation sets, and it was used to train a new QDA model which was then tested with the remaining test set data. The results for this last algorithm are presented in Table 2. The performance of this last model is over 0.95 for sensitivity, recall, and F1 scores for each class. Also, the confusion matrix is added in the results Table 3; this shows which beats were classified correctly and which ones were not.

The present work proposes certain specific features as a way to generalize and reclassify the heartbeats. The results from the algorithms commonly applied give us an insight on the standard methodology followed in the classification of the heartbeats and in the degree of the complexity needed to discern between the classes even though there are some simple linear classifiers able to obtain high percentages in the classification rate.

However, there are much more complex models available with the capacity to learn even nonlinear features which are the state of the art in machine learning algorithms as is seen in Table 4. The classification difficulty varies depending on the kind of classes to detect; anyway most of them have already high percentages in their classification rates. The main goal of our work is not about the reaching of the 100 percent in the classification rate or to achieve better performance than the related works, but instead demonstrating that simple features can have unambiguous boundaries and that such features can reach a probability distribution which

TABLE 4: Comparisons between related works.

Work	Year	Comparison with other works					
		Features	Classifier	Classes	Acc	Se	PPV
Nazarahari et al. [8]	2015	Wavelet + distances measures	Multilayer perception	Normal, PVC, APC, paced, LBBB, RBBB	97.51	—	—
Martis et al. [9]	2013	QRS, bispectrum, PCA	SVM NN	N, LBBB, RBBB, APC, VPC	93.48	—	—
Afkhami et al. [10]	2016	RR interval, HOS, GMM	Decision trees, ensemble learns	AAMI, all classification in MIT-BIH	99.7	100	100
Javadi et al. [11]	2013	Wavelet + morpho-logical and temporal features	Mixture of experts, negative correlation learning	N, PVC, other	96.02	92.27	79.4
Kamath [12]	2011	Teager energy functions in time and frequency domains	Neural network	N, LBBB, RBBB, PVC, paced beats	100	100	100
Martis et al. [13]	2013	DWT + PCA + ICA + LDA	SVM, NN, PNN	AAMI	99.28	—	—
Sharma and Ray [14]	2016	Hilbert–Huang transform, statistical features	SVM	N, LBBB, RBBB, PVC, paced, APC	99.51	99.36	100
Banerjee and Mitra [15]	2014	Cross wavelet transform	Heuristic classification	Abnormal versus normal	97.6	97.3	98.8
Oliveira et al. [16]	2016	Dynamic Bayesian networks	Dynamic threshold	PVC versus others	99.88	99	99
Work		FSC, SFE	QDA	NB, PVC, OB	98.3	100	98

will then be helpful in giving insights about the different types of beats that we are trying to classify.

For example, in [8, 9, 12, 13], we use NNs to discriminate between the different types of beats according to their respective application. The multilayer NNs have been a great tool for classification purposes; they are powerful and their deep learning extension (state of the art for complicated problems like image classification and object detection on images) has put them into a powerful place to solve many problems. However, they are called “black box” algorithms since it is very hard to interpret the hyperparameters they learned through the training process. Also, the mixture of experts used in [11] with temporal-frequency domain features from the wavelet transform does not provide further insights of the classification boundaries which could be extrapolated into a medical interpretation. Although decision trees (applied in [10]) are known for being algorithms used in Business Applications for their great intuitive design and modeling, these along with the features extracted from HOS and GMM may represent a difficult task to find some interoperability; and Ensemble learning makes it even harder.

The SVM was considered state of the art for classification in its time, the kernel approach made suitable for tasks very hard to accomplish, and it is widely used nowadays as in [12, 14] but represents the same lack of interpretability for being a discriminant model along with the complicated feature extraction (Bispectrum, PCA transformations, and temporal-frequency coefficients of Hilbert-Huang transformation). Finally heuristic classification and dynamic threshold in [15, 16] works, respectively, depend on medical expertise to tune the decision. We believe that all these methods and approaches are suitable for the classification purpose, but we also believe that they overkill the issue. In our method, for all this, we get rid of all the unnecessary extra complexity, and we use simple Bayesian models adapted to

explore and analyze the data using our proposed features for these classifications achieving very good results as shown.

The independence assumption, the linear and quadratic boundaries from the Naïve Bayes, the discriminant linear and quadratic classifiers, respectively, and the results from these classifiers give us an insight of the separation between classes with the features that we proposed. These algorithms make evident the advantage that there are no other hyperparameters to tune as in logistic regressions, in support vector machines, in neural networks, and so on. The results are showing us that the boundary among classes is not complicated for a linear classifier using the features proposed. The finding of promising algorithm candidates and methodologies to classify ventricular heartbeats, as well as normal heartbeats and other types of beats, can lead to better treatment and diagnosis of heart issues. Comparing our method with the results of other works puts our approach very close to similar implementations in which complex classifiers were used.

#### 4. Conclusion

We evaluated the Naïve Bayes, LDA, and QDA algorithms to classify ECG beats in normal, PVC, and other kinds. These three models have high F1 scores, with Naïve Bayes presenting 0.86 while LDA and QDA presenting more than 0.95. However, in relation to the last two, the QDA classifier have a higher 0.983 F1 score. This performance results is the reason why we preferred the QDA classifier over the other three models. This chosen classifier was trained with both the training and validation sets, and it was tested with the corresponding data test set giving promising results. The F1 scores for each class were above the 0.95, giving almost the unit value for the PVC class; this was possible due to the fact that the relation of sensitivity and recall for this kind of beats yields better results than those in all the other classes. It can

be proved from the confusion matrix that every PVC beat was correctly classified. The algorithms proposed in this work are, in nature, simple, and given that they are generative these assign a probability distribution to the features which can also give insights about the distinct heartbeats and their behavior. The finding of even more features for the mapping of these heartbeats into a better feature space and for their interpretation in behalf of the medical field is an active area of research.

## Ethical Approval

This research as well as the human data used in it was approved by the Committee of Ethics in Research of the General Hospital of Mexicali BC.

## Conflicts of Interest

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

## Acknowledgments

The authors want to express their gratitude to the National Council of Science and Technology (Conacyt) for its scholarships and materials in support for this research.

## References

- [1] World Health Organization, Fact Sheet No. 317, *Cardiovascular Diseases*, 2015, <http://www.who.int/mediacentre/factsheets/fs317/en/>.
- [2] ACC/AHA/ESC, "ACC/AHA/ESC, guidelines for management of patients with ventricular arrhythmias and the prevention of sudden cardiac death," *Circulation*, vol. 114, pp. 1088–1132, 2006.
- [3] M. S. Thaler, *The Only EKG Book You'll Ever Need*, Wolters Kluwer, Philadelphia, PA, USA, 2015.
- [4] K. Nikus, J. Lhteenmki, P. Lehto, and M. Eskola, "The role of continuous monitoring in a 24/7 telecardiology consultation service—a feasibility study," *Journal of Electrocardiology*, vol. 42, no. 6, pp. 473–480, 2009.
- [5] M. Velic, I. Padavic, and S. Car, "Computer aided ECG analysis—state of the art and upcoming challenges," in *Proceedings of the IEEE Eurocon 2013*, pp. 1778–1784, Zagreb, Croatia, July 2013.
- [6] D. Nabil and F. B. Reguig, "Ectopic beats detection and correction methods: a review," *Biomedical Signal Processing and Control*, vol. 18, pp. 228–244, 2015.
- [7] E. J. D. S. Luz, W. R. Schwartz, C. C. Chvez, and D. Menotti, "ECG-based heartbeat classification for arrhythmia detection: a survey," *Computer Methods and Programs in Biomedicine*, vol. 127, pp. 144–164, 2016.
- [8] M. Nazarahari, S. G. NaminAmir, A. H. D. Markazi, and A. K. Anaraki, "A multi-wavelet optimization approach using similarity measures for electrocardiogram signal classification," *Biomedical Signal Processing and Control*, vol. 20, pp. 142–151, 2015.
- [9] R. J. Martis, U. R. Acharya, K. M. Mandana, A. K. Ray, and C. Chakrabortya, "Cardiac decision making using higher order spectra," *Biomedical Signal Processing and Control*, vol. 8, no. 2, pp. 193–203, 2013.
- [10] R. G. Afkhami, G. Azarnia, and M. A. Tinati, "Cardiac arrhythmia classification using statistical and mixture modeling features of ECG signals," *Pattern Recognition Letters*, vol. 70, pp. 45–51, 2016.
- [11] M. Javadi, S. A. A. Arani, A. Sajedin, and R. Ebrahimpour, "Classification of ECG arrhythmia by a modular neural network based on mixture of experts and negatively correlated learning," *Biomedical Signal Processing and Control*, vol. 8, no. 3, pp. 289–296, 2013.
- [12] C. Kamath, "ECG beat classification using features extracted from teager energy functions in time and frequency domains," *IET Signal Processing*, vol. 5, no. 6, pp. 575–581, 2011.
- [13] R. J. Martis, U. R. Acharya, and L. C. Min, "ECG beat classification using PCA, LDA, ICA and discrete wavelet transform," *Biomedical Signal Processing and Control*, vol. 8, no. 5, pp. 437–448, 2013.
- [14] P. Sharma and K. C. Ray, "Efficient methodology for electrocardiogram beat classification," *IET Signal Processing*, vol. 10, no. 7, pp. 825–832, 2016.
- [15] S. Banerjee and M. Mitra, "Application of cross wavelet transform for ECG pattern analysis and classification," *IEEE Transactions on Instrumentation and Measurement*, vol. 63, no. 2, pp. 326–333, 2013.
- [16] L. S. C. Oliveira, R. V. Andreao, and M. S. Filho, "Bayesian network with decision threshold for heart beat classification," *IEEE Latin America Transactions*, vol. 14, no. 3, pp. 1103–1108, 2016.
- [17] R. Mark, P. Schluter, G. Moody, P. Devlin, and D. Chernoff, "An annotated ECG database for evaluating arrhythmia detectors," in *Proceedings of the 4th Annual Conference IEEE Engineering in Medicine and Biology Society*, pp. 205–210, Philadelphia, PA, USA, September 1982.
- [18] A. L. Goldberger, L. A. N. Amaral, L. Glass et al., "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.
- [19] T. Hastie, R. Tibshirani, and J. Friedman, "The elements of statistical learning," in *Springer Series in Statistics*, Springer, Berlin, Germany, 2009.
- [20] P. J. G. Lisbora, I. O. Ellis, A. R. Green, F. Ambrogi, and M. B. Dias, "Cluster-based visualization with scatter matrices," *Pattern Recognition Letters*, vol. 29, no. 13, pp. 1814–1823, 2008.
- [21] M. M. Casas, R. L. Avitia, M. A. Reyna, and J. A. Cárdenas-Haro, "Evaluation of three machine learning algorithms as classifiers of premature ventricular contractions on ECG beats," in *Proceedings of the 2016 Global Medical Engineering Physics Exchanges/Pan American Health Care Exchanges (GMEPE/PAHCE)*, pp. 1–6, Madrid, Spain, April 2016.

## Research Article

# Predicting Interactions between Virus and Host Proteins Using Repeat Patterns and Composition of Amino Acids

Saud Alguwaizani,<sup>1</sup> Byungkyu Park,<sup>1</sup> Xiang Zhou,<sup>1</sup> De-Shuang Huang <sup>2</sup>  
and Kyungsook Han <sup>1</sup>

<sup>1</sup>Department of Computer Engineering, Inha University, Incheon 22212, Republic of Korea

<sup>2</sup>School of Electronics and Information Engineering, Tongji University, Shanghai 201804, China

Correspondence should be addressed to Kyungsook Han; [khan@inha.ac.kr](mailto:khan@inha.ac.kr)

Received 14 November 2017; Revised 27 March 2018; Accepted 17 April 2018; Published 10 May 2018

Academic Editor: Xiaofeng Yang

Copyright © 2018 Saud Alguwaizani 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.

Previous methods for predicting protein-protein interactions (PPIs) were mainly focused on PPIs within a single species, but PPIs across different species have recently emerged as an important issue in some areas such as viral infection. The primary focus of this study is to predict PPIs between virus and its targeted host, which are involved in viral infection. We developed a general method that predicts interactions between virus and host proteins using the repeat patterns and composition of amino acids. In independent testing of the method with PPIs of new viruses and hosts, it showed a high performance comparable to the best performance of other methods for single virus-host PPIs. In comparison of our method with others using same datasets, our method outperformed the others. The repeat patterns and composition of amino acids are simple, yet powerful features for predicting virus-host PPIs. The method developed in this study will help in finding new virus-host PPIs for which little information is available.

## 1. Introduction

Viral infection involves a large number of protein-protein interactions (PPIs) between virus and its targeted host. These interactions range from the initial binding of viral coat proteins to host membrane receptor to hijack the host transcription machinery by virus proteins. Various viral diseases are caused by infection with pathogenic viruses. For instance, Ebola virus disease is a highly contagious and fatal disease caused by infection with Ebola virus. During the 2014 Ebola epidemic, the world witnessed over 28,000 cases and over 11,000 deaths [1]. So far, there is no specific vaccine or effective treatment for Ebola virus disease [2]. Despite the increased number of known virus-host PPIs, viral infection mechanism is not fully understood. Thus, identifying interactions between virus proteins and host proteins helps understand the mechanism of viral infection and develop treatments and vaccines.

So far, many computational methods have been developed to predict PPIs. However, most of these methods predict PPIs within a single species and cannot be used to

predict PPIs between different species because they do not distinguish interactions between proteins of the same species from those of different species. Recently, a few computational methods have been developed to predict virus-host PPIs using machine learning methods. For instance, a homology-based method [3] predicts PPIs between *H. sapiens* and *M. tuberculosis* H37Rv. Support vector machine (SVM) models developed by Cui et al. [4] and Kim et al. [5] predicted PPIs between human and two types of viruses (hepatitis C virus and human papillomavirus). However, these methods are intended for PPIs between virus of a single type and host of a single type. Recent computational methods developed for predicting virus-host PPIs [6–8] are also limited to PPIs between human and the human immunodeficiency virus 1 (HIV-1) and cannot predict PPIs of new viruses or new hosts which have no known PPIs to the methods. A recent SVM model called DeNovo can exceptionally predict PPIs of new viruses with a shared host [9].

In this paper, we present a new method for predicting virus-host PPIs, which is applicable to new viruses or hosts using amino acid repeat patterns and composition. Proteins

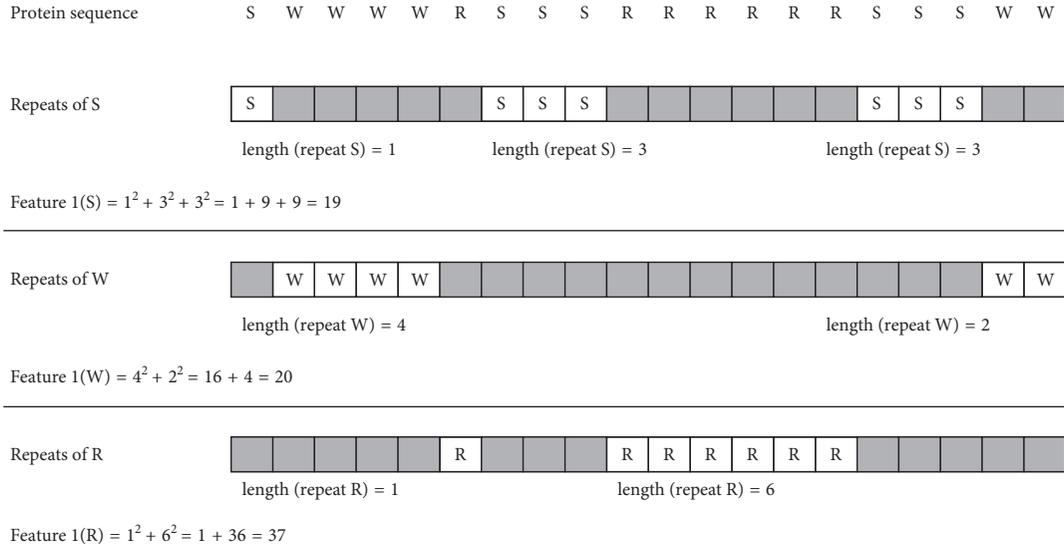


FIGURE 1: Example of computing feature 1 (F1) of amino acid repeats. F1 is the sum of squared length of single amino acid repeats (SARs) in the whole protein sequence. In this example,  $F1(\text{repeats of S}) = 1^2 + 3^2 + 3^2 = 19$ ,  $F1(\text{repeats of W}) = 4^2 + 2^2 = 20$ , and  $F1(\text{repeats of R}) = 1^2 + 6^2 = 37$ .

in a variety of species contain significant amino acid repeats, with more abundance of repeats in eukaryotic proteins than in prokaryotic proteins [10, 11]. It has been found that proteins with a large number of amino acid repeats have a greater number of interacting partners compared to those without [12]. Experimental results of our method show that the repeat patterns and local composition of amino acids are simple, yet powerful features for predicting virus-host PPIs. The rest of this paper discusses the details of the method and its experimental results.

## 2. Materials and Methods

**2.1. Features and Representation.** Proteins are of different lengths and have different amino acid compositions. Many features of proteins have been used to predict PPIs from protein sequences. In this study, we represent a virus-host PPI by three features (F1, F2, and F3):

F1: sum of squared length of single amino acid repeats (SARs) in the entire protein sequence

F2: maximum of the sum of squared length of SARs in a window of 6 residues

F3: composition of amino acids in 5 partitions of the protein sequence

F1, which is the sum of squared length of SARs in the protein sequence, is defined by (1). Since SAR of length 1 is also included in F1, the F1 score reflects global composition of amino acids as well as amino acid repeats. Figure 1 shows an example of how we compute F1.

$$F1(\text{SAR}) = \sum_{\text{SAR} \in \text{sequence}} \text{length}(\text{SAR})^2. \quad (1)$$

Feature F2 is defined by (2). It appears to be similar to F1, but there are two differences: (1) for F2, the sum of squared

length of SARs is computed for every window of size 6 instead of a whole protein sequence, and (2) the maximum of the sum of squared length of SARs in a window is selected for F2. For example, a protein sequence `SWWWWRSSRRRRRRSSSWW` has 15 possible windows of size 6, as shown in Figure 2. For each amino acid, we compute its F2 score by selecting the maximum of the sum of squared length of the SAR in a window of size 6:

$$F2(\text{SAR}) = \max_{\text{window} \in \text{sequence}} \sum_{\text{SAR} \in \text{window}} \text{length}(\text{SAR})^2. \quad (2)$$

The reason that we use a window of size 6 for F2 is because a window larger than 6 residues generates a same score for different repeat patterns. For example, with a window of size 7, we may obtain a same value of F2 even for different patterns of single amino acid repeats, whereas with a window of size 6, we obtain all different values of F2 for different patterns of single amino acid repeats (Figure 3).

While feature F1 represents the repeat patterns and global composition of amino acids in the whole protein sequence, feature F3 represents the local composition of amino acids. For feature F3, we partition a protein sequence into 5 segments of equal length except the last one and compute the composition of amino acids in each of the 5 segments. Since the three features, F1, F2, and F3, are computed for each amino acid, every pair of virus and host proteins is represented in a feature vector with 280 elements (140 for a virus protein and 140 for a host protein).

**2.2. Datasets of Virus-Host PPIs.** We constructed several datasets of virus-host PPIs to examine the applicability of the prediction method to new viruses and hosts. The first training dataset consists of PPIs of human with positive-sense single-stranded RNA (+ssRNA) viruses except hepatitis C virus (HCV) and severe acute respiratory syndrome

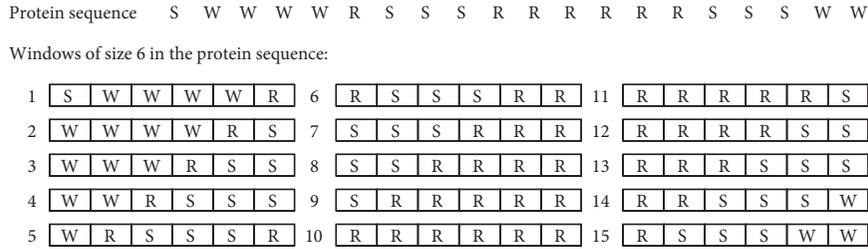


FIGURE 2: Example of computing feature 2 (F2) of amino acid repeats. F2 is the maximum value of the sum of squared length of single amino acid repeats in a window of size six. The maximum repeat size of amino acid S is 3, which is observed in the windows starting at 4, 5, 6, 7, 13, 14, and 15. So, F2 (repeats of S) = 3<sup>2</sup> = 9. The maximum repeat size of amino acid W is 4, observed in the windows starting at 1 and 2. F2 (repeats of W) = 4<sup>2</sup> = 16. The maximum repeat size of amino acid R is 6, observed in the window starting at 10. F2 (repeats of R) = 6<sup>2</sup> = 36.

Case	Window of size 6						Equation for feature #2	Value
1							$0^2 = 0$	0
2	.						$1^2 = 1$	1
3	.	.					$1^2 + 1^2 = 2$	2
4	.	.	.				$1^2 + 1^2 + 1^2 = 3$	3
5	.	.	.	.			$2^2 = 4$	4
6	.	.	.	.	.		$2^2 + 1^2 = 5$	5
7	.	.	.	.	.	.	$2^2 + 1^2 + 1^2 = 6$	6
8	.	.	.	.	.	.	$2^2 + 2^2 = 8$	8
9	.	.	.	.	.	.	$3^2 = 9$	9
10	.	.	.	.	.	.	$3^2 + 1^2 = 10$	10
11	.	.	.	.	.	.	$3^2 + 2^2 = 13$	13
12	.	.	.	.	.	.	$4^2 = 16$	16
13	.	.	.	.	.	.	$4^2 + 1^2 = 17$	17
14	.	.	.	.	.	.	$5^2 = 25$	25
15	.	.	.	.	.	.	$6^2 = 36$	36

Case	Window of size 7						Equation for feature #2	Value
1	.	.	.	.	.	.	$1^2 + 1^2 + 1^2 + 1^2 = 4$	4
2	.	.	.	.	.	.	$2^2 = 4$	4

FIGURE 3: Values of feature 2 for windows of six and seven residues. With a window of size 6, different patterns of single amino acid repeats lead to 15 different values for feature 2. With a window of size 7, different patterns of single amino acid repeats can lead to a same value for feature 2, as shown in this example.

(SARS) virus. The SVM model trained with the training dataset was tested on PPIs of five new viruses: HCV, SARS virus, influenza A (H1N1) virus, human papillomavirus (HPV-16), and human immunodeficiency virus HIV-1. Both HCV and SARS are positive-sense single-stranded RNA (+ssRNA) viruses, but the remaining three viruses are of different type. H1N1 virus is a negative-sense single-stranded RNA (-ssRNA) virus, HPV-16 is a double-stranded DNA (dsDNA) virus, and HIV-1 is a retrovirus.

The second training dataset is composed of PPIs between human and +ssRNA viruses, including HCV and SARS virus. The SVM model trained on the second training dataset was tested on PPIs of new hosts: *Mus musculus*, *Bos taurus*, *Rattus norvegicus*, *Sus scrofa*, and *Escherichia coli* K-12.

Data of virus-host PPIs were collected from IntAct [13] and VirusMentha [14]. But PPIs of HCV with human were

obtained from the Hepatitis C Virus Protein Interaction Database (HCVpro) [15] because HCVpro has more human-HCV PPIs than IntAct. The sequences of the proteins involved in the virus-host PPIs were obtained from the UniProt database [16].

The training and test datasets constructed in our study can be summarized as follows.

1. Training (TR) and Test (TS) Datasets for Assessing the Applicability of the Prediction Model to New Viruses

TR1: 638 PPIs between 499 human proteins and 25 +ssRNA virus proteins

TS1: 515 PPIs between 423 human proteins and 11 HCV proteins

TS2: 30 PPIs between 27 human proteins and 12 SARS virus proteins

TS3: 377 PPIs between 307 human proteins and 10 H1N1 virus proteins

TS4: 319 PPIs between 298 human proteins and 11 HPV-16 proteins

TS5: 1,578 PPIs between 1,056 human proteins and 46 HIV-1 proteins

## 2. Training (TR) and Test (TS) Datasets for Assessing the Applicability of the Prediction Model to New Hosts

TR2: 689 PPIs between 522 human proteins and 35 +ssRNA virus proteins

TS6: 191 PPIs between 141 *Mus musculus* proteins and 116 virus proteins

TS7: 125 PPIs between 87 *Bos taurus* proteins and 34 virus proteins

TS8: 86 PPIs between 79 *Rattus norvegicus* proteins and 24 virus proteins

TS9: 57 PPIs between 38 *Sus scrofa* proteins and 10 virus proteins

TS10: 78 PPIs between 64 *Escherichia coli* K-12 proteins and 27 virus proteins

Machine learning-based approaches to PPI prediction require both positive and negative PPI data, but negative data are not available in databases. Constructing a negative dataset of PPIs is not straightforward because there is no experimentally verified noninteracting pair [17]. Eid et al. [9], for example, used negative sampling for their negative dataset. In our study, we constructed a negative dataset with human proteins whose sequence similarity is lower than 40% to any human protein in the positive dataset by running CD-HIT [18]. Our negative dataset includes 2,819 interactions between 90 virus proteins and 2,819 human proteins. The training and test datasets constructed in this study are available in Additional files 1 and 2.

**2.3. Prediction Models of Virus-Host PPIs.** We built several support vector machine (SVM) models using LIBSVM [19] to evaluate our approach. The radial basis function (RBF) was used as a kernel of the SVM models, and the best values of parameters  $C$  and  $\gamma$  were obtained by running the grid search of LIBSVM on training datasets. Unless specified otherwise, the results shown in this paper were obtained with  $C=2$  and  $\gamma=0.5$ . The SVM models take a pair of virus and host protein sequences as input. As output, the SVM models classify whether or not the virus protein interacts with the host protein. The SVM models and supporting data are available at <http://www.bclab.inha.ac.kr/VHPPI>.

## 3. Results and Discussion

**3.1. Performance Measures.** The performance of the SVM models was evaluated by several measures: sensitivity (Sn), specificity (Sp), accuracy (Acc), positive predictive value (PPV), negative predictive value (NPV), and Matthews correlation coefficient (MCC), which are defined by the following equations:

$$Sn = \frac{TP}{TP + FN} \quad (3)$$

$$Sp = \frac{TN}{TN + FP} \quad (4)$$

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \quad (5)$$

$$PPV = \frac{TP}{TP + FP} \quad (6)$$

$$NPV = \frac{TN}{TN + FN} \quad (7)$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (8)$$

In (3)–(8), true positives (TP) are host proteins that are correctly predicted as interacting with a virus protein. True negatives (TN) are noninteracting host proteins that are correctly predicted as noninteracting with a virus protein. False positives (FP) are noninteracting host proteins that are incorrectly predicted as interacting with a virus protein. False negatives (FN) are interacting host proteins that are incorrectly predicted as noninteracting with a virus protein.

**3.2. Results of Cross Validation.** We performed 10-fold cross validation of the SVM model with several datasets which contain different ratios (1:1, 1:2, and 1:3) of positive to negative PPIs between +ssRNA viruses and human. As shown in Table 1, the best performance of the SVM model was observed in the balanced dataset with 1:1 ratio of positive to negative data. As expected, running the SVM model on unbalanced datasets resulted in lower performances than running it on the balanced dataset with 1:1 ratio of positive to negative data. Datasets are available in Additional file 3.

We also examined the contribution of the features to the prediction performance of the SVM model. Table 2 shows the results of using different combinations of features in 10-fold cross validation of the SVM model with the 1:1 dataset of Table 1. Among the single features, F3, which is the local composition of amino acids, was the best in all performance measures. With F3 alone, the SVM model achieved an accuracy above 92% and an MCC above 0.86, indicating that F3 is a very powerful feature in predicting virus-host PPIs. The best performance of the SVM model was observed when F1 and F3 were used. We also examined this work with different combinations of features. We used double amino acid repeats (DARs) for F1 and F2 instead of single amino acid repeats (SARs), but here for F2, we used a window size of 10 residues not 6 residues because we are working with DAR, so a window size of 10 residues is the biggest available window size that obtain a different value for every double amino acid repeat in it, but a window size of 6 residues does the same thing for the single amino acid repeat.

For features F1 and F2, we tried both single amino acid repeats (SARs) and double amino acid repeats (DARs) along

TABLE 1: Results of 10-fold cross validation of SVM model on 1,072 PPIs between 36 RNA viruses and 812 human proteins with different ratios of positive to negative instances.

P:N	Dataset	Sn (%)	Sp (%)	Acc (%)	PPV (%)	NPV (%)	MCC	AUC
1:1	1	88.24	97.34	92.79	97.07	89.22	0.859	0.963
	2	81.03	94.36	87.7	93.49	83.26	0.761	0.931
	3	77.74	94.04	85.89	92.88	80.86	0.728	0.926
	mean $\pm$ SD	82.34 $\pm$ 4.39	95.25 $\pm$ 1.49	88.79 $\pm$ 2.92	94.48 $\pm$ 1.85	84.45 $\pm$ 3.51	0.78 $\pm$ 0.06	0.94 $\pm$ 0.02
1:2	1	64.89	97.34	86.52	92.41	84.72	0.693	0.893
	2	58.31	97.57	84.48	92.31	82.4	0.646	0.886
	3	63.64	96.08	85.27	89.04	84.09	0.661	0.891
	mean $\pm$ SD	62.28 $\pm$ 2.85	97 $\pm$ 0.65	85.42 $\pm$ 0.84	91.25 $\pm$ 1.57	83.74 $\pm$ 0.98	0.67 $\pm$ 0.02	0.89 $\pm$ 0.00
1:3	1	46.24	98.28	85.27	89.94	84.58	0.58	0.850
	2	46.87	98.59	85.66	91.72	84.77	0.59	0.863
	3	49.37	97.28	85.31	85.83	85.22	0.576	0.858
	mean $\pm$ SD	47.49 $\pm$ 1.35	98.05 $\pm$ 0.56	85.41 $\pm$ 0.18	89.16 $\pm$ 2.47	84.86 $\pm$ 0.27	0.58 $\pm$ 0.01	0.86 $\pm$ 0.01

Sn: sensitivity, Sp: specificity, Acc: accuracy, PPV: positive predictive value, NPV: negative predictive value, MCC: Matthews correlation coefficient, and AUC: the area under the ROC curve.

TABLE 2: Comparison of different combinations of features in 10-fold cross validation of SVM model.

Features	Sn (%)	Sp (%)	Acc (%)	PPV (%)	NPV (%)	MCC	AUC
F1	81.66	97.02	89.34	96.48	84.10	0.796	0.916
F2	69.75	85.11	77.43	82.41	73.78	0.555	0.849
F3	87.78	97.81	92.79	97.56	88.89	0.860	0.965
F1 + F2	80.88	95.61	88.24	94.85	83.33	0.773	0.925
F1 + F3	88.56	97.34	92.94	97.08	89.48	0.862	0.966
F2 + F3	87.46	96.87	92.16	96.54	88.54	0.847	0.961
F1 + F2 + F3	88.24	97.34	92.79	97.07	89.22	0.859	0.963

F1: sum of squared length of single amino acid repeats in the entire protein sequence, F2: maximum of the sum of squared length of single amino acid repeats in a window of 6 residues, F3: composition of amino acids in 5 partitions of the protein sequence, Sn: sensitivity, Sp: specificity, Acc: accuracy, PPV: positive predictive value, NPV: negative predictive value, MCC: Matthews correlation coefficient, and AUC: the area under the ROC curve.

TABLE 3: Results of 10-fold cross validation of SVM model on different combinations of the three features we used in our method.

F1 and F2	F3	Sn (%)	Sp (%)	Acc (%)	PPV (%)	NPV (%)	MCC	AUC
SAR	5 partitions	88.24	97.34	92.79	97.07	89.22	0.859	0.963
SAR	7 partitions	88.24	97.96	93.10	97.74	89.29	0.866	0.965
SAR	9 partitions	89.19	96.08	92.63	95.79	89.88	0.855	0.962
DAR	5 partitions	84.80	94.51	89.66	93.92	86.14	0.797	0.937
DAR	7 partitions	85.42	94.51	89.97	93.97	86.64	0.803	0.938
DAR	9 partitions	85.27	94.20	89.73	93.63	86.47	0.798	0.940

SAR: single amino acid repeats for F1 and F2, DAR: double amino acid repeats for F1 and F2, Sn: sensitivity, Sp: specificity, Acc: accuracy, PPV: positive predictive value, NPV: negative predictive value, MCC: Matthews correlation coefficient, and AUC: the area under the ROC curve.

with different partitions of a protein sequence. As shown in Table 3, SAR resulted in a better performance than DAR.

For feature F3, we tried several different partitions of a protein sequence in several datasets. Table 4 shows the performance of our SVM model in three different datasets of virus-host PPIs. All the results shown in Table 4 were obtained by using SAR for features F1 and F2, but with different partitions for feature F3. On average, partitioning a protein sequence into 5 segments showed the best performance in all performance measures except sensitivity. In addition to the performance gain, partitioning a protein sequence into 5 segments is more advantageous than 7 or 9 segments with respect to the size of a feature vector that represents the sequence. When we partition a protein sequence into 5 segments, every pair of virus and host proteins is encoded in a feature vector with 280 elements (20 elements

for F1, 20 elements for F2, and  $20 \times 5 = 100$  elements for F3 for each of the virus and host proteins). If we partition a protein sequence into 7 or 9 partitions, a feature vector will require 360 elements (20 elements for F1, 20 elements for F2, and  $20 \times 7 = 140$  elements for F3 for each of the virus and host proteins) or 440 elements (20 elements for F1, 20 elements for F2, and  $20 \times 9 = 180$  elements for F3 for each of the virus and host proteins). However, the larger feature vectors did not result in performance improvement in predicting virus-host PPIs.

*3.3. Results of Independent Testing on PPIs of New Viruses.* As discussed earlier, we trained the SVM model with the training dataset TR1 consisting of PPIs of human with +ssRNA viruses except hepatitis C virus (HCV) and SARS

TABLE 4: Results of testing our SVM model with different partitions of a protein sequence on three datasets.

<i>Our dataset</i>							
F3	Sn (%)	Sp (%)	Acc (%)	PPV (%)	NPV (%)	MCC	AUC
5 partitions	88.24	97.34	92.79	97.07	89.22	0.859	0.963
7 partitions	88.24	97.96	93.10	97.74	89.29	0.866	0.965
9 partitions	89.19	96.08	92.63	95.79	89.88	0.855	0.962
<i>DeNovo dataset</i>							
F3	Sn (%)	Sp (%)	Acc (%)	PPV (%)	NPV (%)	MCC	AUC
5 partitions	86.35	86.59	86.47	86.56	86.39	0.729	0.926
7 partitions	83.60	81.18	82.41	82.30	82.54	0.648	0.907
9 partitions	84.27	79.53	81.95	81.17	82.84	0.639	0.902
<i>Barman dataset</i>							
F3	Sn (%)	Sp (%)	Acc (%)	PPV (%)	NPV (%)	MCC	AUC
5 partitions	73.72	83.48	78.60	81.69	76.06	0.575	0.847
7 partitions	78.55	78.55	78.55	78.55	78.55	0.571	0.858
9 partitions	78.16	79.81	78.99	79.47	78.52	0.580	0.860
<i>Average of the above three results</i>							
F3	Sn (%)	Sp (%)	Acc (%)	PPV (%)	NPV (%)	MCC	AUC
5 partitions	82.77	89.14	85.95	88.44	83.89	0.721	0.912
7 partitions	83.46	85.90	84.69	86.20	83.46	0.695	0.910
9 partitions	83.87	85.14	84.52	85.48	83.75	0.691	0.908

All the results were obtained by commonly using SAR for features F1 and F2.

TABLE 5: Training (TR) and test (TS) datasets for assessing the applicability of the SVM model to new viruses and to new hosts. The average sequence similarity between proteins in TR and those in TS was analyzed using EMBOSS Needle tool [20].

Proteins in training datasets	Target proteins in test datasets	Average sequence similarity (%)
25 virus proteins in TR1	11 HCV proteins in TS1	5.03
	12 SARS virus proteins in TS2	5.20
	10 H1N1 virus proteins in TS3	5.03
	11 HPV-16 proteins in TS4	3.12
	46 HIV-1 proteins in TS5	3.56
522 human proteins in TR2	141 <i>Mus musculus</i> proteins in TS6	9.20
	87 <i>Bos taurus</i> proteins in TS7	9.07
	79 <i>Rattus norvegicus</i> proteins in TS8	9.76
	38 <i>Sus scrofa</i> proteins in TS9	8.70
	64 <i>Escherichia coli</i> K-12 proteins in TS10	8.04

virus and tested it on PPIs of new viruses which were not used in training the model. The test datasets include PPIs of five viruses (HCV, SARS virus, H1N1 virus, HPV-16, and HIV-1) with human. H1N1 virus is a negative-sense single-stranded RNA (-ssRNA) virus, and HPV-16 is a double-stranded RNA (dsDNA) virus. HIV-1 is a retrovirus, which is a +ssRNA virus with a DNA intermediate.

In addition to species difference, we also analyzed the sequence similarity between the training datasets and test datasets using EMBOSS Needle tool [20] to assess the independence of the test data from the training data. As shown in Table 5, target virus proteins in the test datasets showed a very low average sequence similarity in the range (3.12% to

TABLE 6: Results of independent testing our SVM on PPIs of new viruses with human.

Virus	Sn (%)	Sp (%)	Acc (%)	PPV (%)	NPV (%)	MCC	AUC
HCV	94.37	52.04	73.20	66.30	90.24	0.512	0.925
SARS virus	96.67	73.33	85.00	78.38	95.65	0.720	0.970
H1N1 virus	90.72	67.90	79.31	73.87	87.97	0.602	0.938
HPV-16	81.82	94.04	87.93	93.21	83.80	0.764	0.938
HIV-1	87.83	64.64	76.24	71.30	84.16	0.539	0.882
Average	90.28	70.39	80.34	76.61	88.36	0.627	0.930

Sn: sensitivity, Sp: specificity, Acc: accuracy, PPV: positive predictive value, NPV: negative predictive value, MCC: Matthews correlation coefficient, and AUC: the area under the ROC curve.

5.20%) to the virus proteins in the training dataset (see Additional file 4 for the similarity of every sequence pair between the training and test datasets).

Table 6 shows the results of testing the prediction model on 5 independent datasets of PPIs of new viruses. Despite such a low sequence similarity and species difference, the SVM model showed a high performance in independent testing. In particular, the SVM model showed a higher sensitivity (94.37% and 96.67%) for HCV and SARS virus, which are +ssRNA viruses. It is interesting to note that HPV-16, which is a dsDNA virus, showed the highest specificity of 94.04% and accuracy of 87.93%. Figure 4 shows the ROC curves of independent testing of the SVM model on PPIs of five new viruses.

### 3.4. Results of Independent Testing on PPIs of New Hosts.

In order to examine the applicability of the SVM model to new hosts, we tested it on PPIs of viruses with new hosts, which were not used in training the model. As described earlier, the model trained with PPIs of human with +ssRNA viruses was tested on PPIs of five new hosts (*Mus musculus*, *Bos taurus*, *Rattus norvegicus*, *Sus scrofa*, and *Escherichia coli* K-12) with the viruses. As shown earlier in Table 5, the average sequence similarity of the human proteins in the training dataset to the new hosts is low, ranging between 8.04% and 9.76%. Despite the low sequence similarity and species difference, testing the model on PPIs of new hosts showed a relatively good performance (Table 7). Figure 5 shows the ROC curves of independent testing of the SVM model on PPIs of five new hosts.

It is interesting to note that proteins of new hosts have a higher average sequence similarity to those in training datasets than proteins of new viruses, but the SVM model showed a lower performance for new hosts. This can be explained by the number of partner proteins of the target proteins shared by training and test datasets. As shown in Table 8, the number of common proteins between the test datasets for new viruses (TS1-TS5) and their training dataset TR1 is larger than the number of common proteins between the test datasets for new hosts (TS6-TS10) and their training dataset TR2. Thus, the SVM model showed a better performance for new viruses than for new hosts.

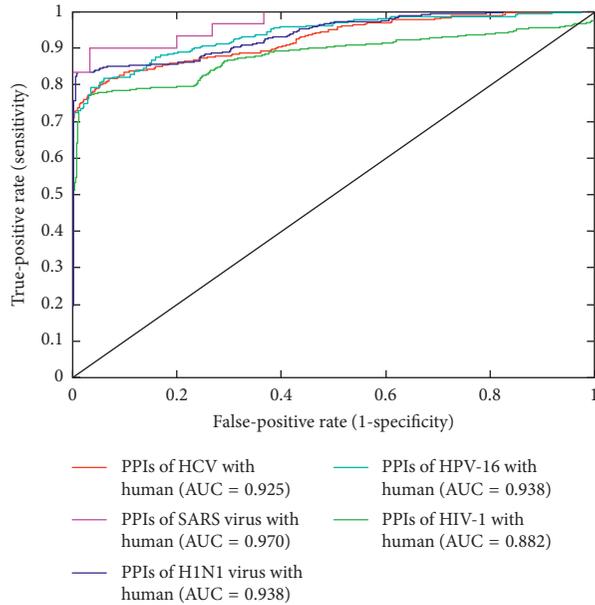


FIGURE 4: ROC curves of independent testing of the SVM model on PPIs of new viruses with human. The SVM model showed the largest area under the ROC curve (AUC) of 0.970 for the PPIs of the SARS virus with human.

TABLE 7: Results of independent testing our SVM on PPIs of new hosts with viruses.

Host	Sn (%)	Sp (%)	Acc (%)	PPV (%)	NPV (%)	MCC	AUC
<i>Mus musculus</i>	85.86	61.78	73.82	69.20	81.38	0.491	0.887
<i>Bos taurus</i>	98.40	27.20	62.80	57.48	94.44	0.365	0.926
<i>Rattus norvegicus</i>	91.86	27.90	59.88	56.03	77.42	0.257	0.828
<i>Sus scrofa</i>	100.00	5.26	52.63	51.35	100.00	0.164	0.952
<i>Escherichia coli K-12</i>	94.87	91.03	92.95	91.36	94.67	0.860	0.959
Average	92.02	54.23	73.13	67.80	86.86	0.501	0.911

Sn: sensitivity, Sp: specificity, Acc: accuracy, PPV: positive predictive value, NPV: negative predictive value, MCC: Matthews correlation coefficient, and AUC: the area under the ROC curve.

These results corroborate the known problem with pair-input methods, which was first reported by Park and Marcotte [21]. According to their study [21], prediction methods that operate on pairs of objects such as PPIs perform much better for test pairs that share components with a training set than for those that do not. Thus, our prediction model showed a better performance in testing for new viruses which share more partner proteins (i.e., host proteins) with training datasets than in testing for new hosts which share fewer partner proteins (i.e., virus proteins) with training datasets.

**3.5. Comparison to Other Methods.** For a comparative purpose, we ran our SVM model on the datasets of two other methods for virus-host PPIs: Barman's method [22] and

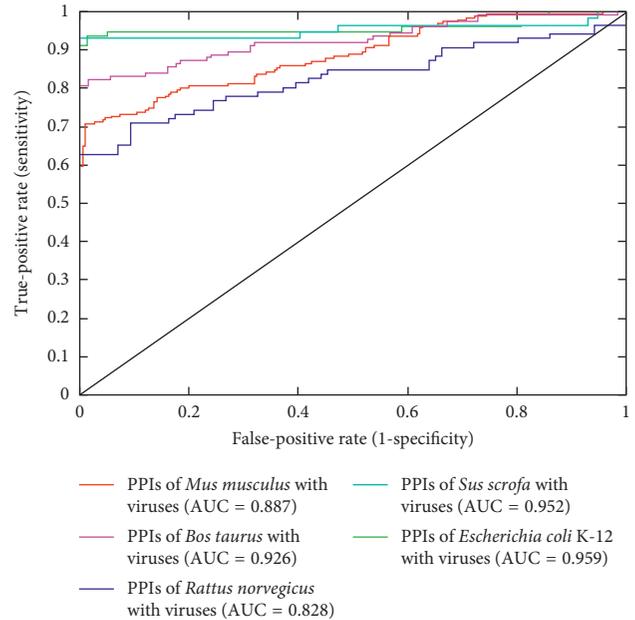


FIGURE 5: ROC curves of independent testing of the SVM model on PPIs of new hosts with viruses. The SVM model showed the largest area under the ROC curve (AUC) of 0.959 for the PPIs of *E. coli* K-12 with viruses.

DeNovo [9]. In Barman's study [22], three machine learning methods (SVM, Naive Bayes, and Random Forest) were used to predict virus-host PPIs using several features such as domain-domain association in interacting protein pairs and composition of methionine, serine, and valine in virus proteins. In a 5-fold cross validation with virus-host PPIs from VirusMINT [23], their Random Forest (RF) and SVM showed a better performance than Naive Bayes. Thus, we tested our SVM model on the same dataset used in Barman's study, which contains 1,035 positive and 1,035 negative interactions between 160 virus proteins of 65 types and 667 human proteins. As shown in Table 9, our SVM model outperformed Barman's SVM model in all performance measures and our SVM model outperformed Barman's RF model in all performance measures except specificity and PPV. The dataset used for comparison of our SVM model with Barman's SVM and Random Forest models is available in Additional file 5.

For comparison with DeNovo's SVM model, we tested our SVM model on DeNovo's SLiM testing set, which contains 425 positive and 425 negative PPIs (Supplementary file S12 used in DeNovo's study ST6). As shown in Table 10, our SVM model was better than DeNovo in all performance measures (sensitivity of 86%, specificity of 87%, and accuracy of 86%). The dataset used for comparison of our SVM model with DeNovo is available in Additional file 6.

## 4. Conclusions

Amino acid repeats are prevalent in a variety of proteins but are rarely used in predicting PPIs. We developed a new method that predicts potential interactions between virus

TABLE 8: The number of host proteins shared by training (TR) and test (TS) datasets used for assessing the applicability of the SVM model to new viruses and to new hosts.

Dataset	TR1	TS1	TR1	TS2	TR1	TS3	TR1	TS4	TR1	TS5
#PPIs	638	515	638	30	638	377	638	319	638	1578
#Virus proteins	25	11	25	12	25	10	25	11	25	46
#Host proteins	499	424	499	27	499	307	499	298	499	1056
#Host proteins common to TR and TS	63 (14.9%)		5 (18.5%)		68 (22.1%)		22 (7.4%)		122 (11.6%)	
Dataset	TR2	TS6	TR2	TS7	TR2	TS8	TR2	TS9	TR2	TS10
#PPIs	689	191	689	125	689	86	689	57	689	78
#Virus proteins	35	116	35	34	35	24	35	10	35	27
#Host proteins	522	141	522	87	522	79	522	38	522	64
#Virus proteins common to TR and TS	9 (7.8%)		1 (2.9%)		4 (16.7%)		0 (0.0%)		0 (0.0%)	

The numbers in parentheses represent the proportion of common proteins to proteins in test datasets.

TABLE 9: Results of 5-fold cross validation of our SVM and Barman's SVM [22] with Barman's dataset of 1,035 positive and 1,035 negative PPIs.

Method	Sn (%)	Sp (%)	Acc (%)	PPV (%)	NPV (%)	MCC	AUC	F1 (%)
Our SVM	73.72	83.48	78.60	81.69	76.06	0.575	0.847	77.50
Barman's SVM	67.00	74.00	71.00	72.00	—	0.440	0.730	69.41
Barman's Random Forest	55.66	89.08	72.41	82.26	—	0.480	0.760	66.39

Sn: sensitivity, Sp: specificity, Acc: accuracy, PPV: positive predictive value, NPV: negative predictive value, MCC: Matthews correlation coefficient, AUC: the area under the ROC curve,  $F1 = 2 \times (SN \times PPV) / (SN + PPV)$ , and "—": not available.

TABLE 10: Results of testing our SVM and DeNovo's SVM [9] on DeNovo's dataset of 425 positive and 425 negative PPIs.

Method	Sn (%)	Sp (%)	ACC (%)	PPV (%)	NPV (%)	MCC	AUC
Our SVM	86.35	86.59	86.47	86.56	86.39	0.729	0.926
DeNovo's SVM	80.71	83.06	81.90	—	—	—	—

Sn: sensitivity, Sp: specificity, Acc: accuracy, PPV: positive predictive value, NPV: negative predictive value, MCC: Matthews correlation coefficient, AUC: the area under the ROC curve, and "—": not available.

and host proteins using global and local compositions of amino acids as well as amino acid repeat patterns.

We tested the prediction model on independent datasets of virus-host PPIs, which were not used in training the model and have a very low sequence similarity to any protein in training datasets of the model. Despite a low sequence similarity between proteins in training datasets and target proteins in test datasets, the prediction model showed a high performance comparable to the best performance of other methods for single virus-host PPIs. In comparison of our method with others using same datasets, our method outperformed the others. Experimental results demonstrate that the repeat patterns and composition of amino acids are simple, yet powerful features for predicting virus-host PPIs. The method can be used to find potential PPIs of new viruses or hosts, for which little information is known.

## Conflicts of Interest

The authors declare no conflicts of interest.

## Acknowledgments

This work was supported by Inha University Research Grant.

## Supplementary Materials

The supporting data and program are available at <http://www.bclab.inha.ac.kr/VHPPI>. Additional file 1: training and test datasets for examining the applicability of the SVM model to new viruses. Additional file 2: training and test datasets for examining the applicability of the SVM model to new hosts. Additional file 3: training datasets with different ratios of positive to negative instances (1 : 1, 1 : 2, and 1 : 3). Additional file 4: the sequence similarity of each protein pair ( $x, y$ ) for  $x$  in the training dataset and  $y$  in the test dataset. Additional file 5: training and test datasets of virus-host PPIs from Barman's study [22]. Additional file 6: training and test datasets of SLiMs from DeNovo [9]. (*Supplementary Materials*)

## References

- [1] M.-A. de La Vega, D. Stein, and G. P. Kobinger, "Ebola virus evolution. Past and present," *PLoS Pathogens*, vol. 11, no. 11, pp. 1–10, 2015.
- [2] D. Gatherer, "The 2014 Ebola virus disease outbreak in West Africa," *Journal of General Virology*, vol. 95, no. 8, pp. 1619–1624, 2014.
- [3] H. Zhou, S. Gao, N. Ninh Nguyen et al., "Stringent homology-based prediction of *H. sapiens-M. tuberculosis*

- H37Rv protein-protein interactions,” *Biology Direct*, vol. 9, no. 1, p. 5, 2014.
- [4] G. Cui, C. Fang, and K. Han, “Prediction of protein-protein interactions between viruses and human by an SVM model,” *BMC Bioinformatics*, vol. 13, no. 7, p. S5, 2012.
- [5] B. Kim, S. Alguwaizani, X. Zhou, D.-S. Huang, B. Park, and K. Han, “An improved method for predicting interactions between virus and human proteins,” *Journal of Bioinformatics and Computational Biology*, vol. 15, no. 1, p. 1650024, 2017.
- [6] B. Barnes, M. Karimloo, A. Schoenrock et al., “Green predicting novel protein-protein interactions between the HIV-1 virus and *Homo sapiens*,” in *Proceedings of the IEEE EMBS International Student Conference (ISC)*, Ottawa, Canada, May 2016.
- [7] A. Becerra, V. A. Bucheli, and P. A. Moreno, “Prediction of virus-host protein-protein interactions mediated by short linear motifs,” *BMC Bioinformatics*, vol. 18, p. 163, 2017.
- [8] S. Ray and A. Sanghamitra Bandyopadhyay, “NMF based approach for integrating multiple data sources to predict HIV-1 human PPIs,” *BMC Bioinformatics*, vol. 17, p. 121, 2016.
- [9] F.-E. Eid, M. ElHefnawi, and L. S. Heath, “DeNovo: virus-host sequence-based protein-protein interaction prediction,” *Bioinformatics*, vol. 32, no. 8, pp. 1144–1150, 2016.
- [10] M. V. Katti, R. Sami-Subbu, P. K. Ranjekar, and V. S. Gupta, “Amino acid repeat patterns in protein sequences: their diversity and structural-functional implications,” *Protein Science*, vol. 9, no. 6, pp. 1203–1209, 2000.
- [11] H. Luo and H. Nijveen, “Understanding and identifying amino acid repeats,” *Briefings in Bioinformatics*, vol. 15, no. 4, pp. 582–591, 2014.
- [12] A. Coletta, J. W. Pinney, D. Y. Weiss Solís, J. Marsh, S. R. Pettifer, and T. K. Attwood, “Low-complexity regions within protein sequences have position dependent roles,” *BMC Systems Biology*, vol. 4, no. 1, p. 43, 2010.
- [13] S. Orchard, M. Ammari, B. Aranda et al., “The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases,” *Nucleic Acids Research*, vol. 42, no. 1, pp. D358–D363, 2014.
- [14] A. Calderone, L. Licata, and G. Cesareni, “VirusMentha: a new resource for virus-host protein interactions,” *Nucleic Acids Research*, vol. 43, no. 1, pp. D588–D592, 2015.
- [15] S. K. Kwofie, U. Schaefer, V. S. Sundararajan, V. B. Bajic, and A. Christoffels, “HCVpro: hepatitis C virus protein interaction database,” *Infection, Genetics and Evolution*, vol. 11, no. 8, pp. 1971–1977, 2011.
- [16] The UniProt Consortium, “UniProt: the universal protein knowledgebase,” *Nucleic Acids Research*, vol. 45, no. 1, pp. D158–D169, 2017.
- [17] E. Nourani, F. Khunjush, and S. Durmuş, “Computational approaches for prediction of pathogen-host protein-protein interactions,” *Frontiers in Microbiology*, vol. 6, p. 94, 2015.
- [18] Y. Huang, B. Niu, Y. Gao, L. Fu, and W. Li, “CD-HIT Suite: a web server for clustering and comparing biological sequences,” *Bioinformatics*, vol. 26, no. 5, pp. 680–682, 2010.
- [19] C.-C. Chang and C.-J. Lin, “LIBSVM: a library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1–27, 2011.
- [20] P. Rice, I. Longden, and A. Bleasby, “EMBOSS: the European Molecular Biology Open Software Suite,” *Trends in Genetics*, vol. 16, no. 6, pp. 276–277, 2000.
- [21] Y. Park and E. M. Marcotte, “A flaw in the typical evaluation scheme for pair-input computational predictions,” *Nature Methods*, vol. 9, no. 12, pp. 1134–1136, 2012.
- [22] R. K. Barman, S. Saha, and S. Das, “Prediction of interactions between viral and host proteins using supervised machine learning methods,” *PLoS One*, vol. 9, no. 11, article e112034, 2014.
- [23] A. Chatr-aryamontri, A. Ceol, D. Peluso et al., “VirusMINT: a viral protein interaction database,” *Nucleic Acids Research*, vol. 37, no. 1, pp. D669–D673, 2009.

## Research Article

# Breast Mass Detection in Digital Mammogram Based on Gestalt Psychology

Hongyu Wang <sup>1</sup>, Jun Feng <sup>1</sup>, Qirong Bu,<sup>1</sup> Feihong Liu,<sup>1</sup> Min Zhang,<sup>2</sup> Yu Ren,<sup>3</sup> and Yi Lv<sup>4</sup>

<sup>1</sup>Department of Information Science and Technology, Northwest University, Xi'an 710127, China

<sup>2</sup>School of Mathematics, Northwest University, Xi'an 710127, China

<sup>3</sup>Department of Breast Surgery, School of Medicine, The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an 710061, China

<sup>4</sup>National Local Joint Engineering Research Center for Precision Surgery and Regenerative Medicine, Xi'an Jiaotong University, Xi'an 710061, China

Correspondence should be addressed to Jun Feng; [fengjun@nwu.edu.cn](mailto:fengjun@nwu.edu.cn)

Received 28 September 2017; Revised 18 January 2018; Accepted 14 March 2018; Published 2 May 2018

Academic Editor: Yong Xia

Copyright © 2018 Hongyu 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.

Inspired by gestalt psychology, we combine human cognitive characteristics with knowledge of radiologists in medical image analysis. In this paper, a novel framework is proposed to detect breast masses in digitized mammograms. It can be divided into three modules: sensation integration, semantic integration, and verification. After analyzing the progress of radiologist's mammography screening, a series of visual rules based on the morphological characteristics of breast masses are presented and quantified by mathematical methods. The framework can be seen as an effective trade-off between bottom-up sensation and top-down recognition methods. This is a new exploratory method for the automatic detection of lesions. The experiments are performed on Mammographic Image Analysis Society (MIAS) and Digital Database for Screening Mammography (DDSM) data sets. The sensitivity reached to 92% at 1.94 false positive per image (FPI) on MIAS and 93.84% at 2.21 FPI on DDSM. Our framework has achieved a better performance compared with other algorithms.

## 1. Introduction

Breast cancer is responsible for 23% of all cancer cases and 14% of cancer-related deaths amongst women worldwide [1]. Mammography is a reliable and trustworthy tool for early detection of breast cancer [2, 3]. Early detection of potential abnormalities could generate a recommendation for further examination [4]. Current progress has shown that computer-aided detection (CAD) systems can assist doctors in finding breast masses from digitized mammograms at an early stage, which greatly improves doctor's working efficiency [5]. Efficient CAD systems have potential to reduce the breast biopsies and release radiologists from heavy workload [6, 7].

However, CAD for breast mass detection is a challenging task due to the varying sizes, shapes, and appearances as well as densities of masses [8, 9]. Conventional methods for breast mass detection mainly rely on the threshold values [10] or mass templates [11] based on various kinds of filter

operators. However, the false-positive results are still very high [12, 13]. The threshold methods based on gray-level images or various filtered images consider only the simple brightness of masses. Although there are ongoing research studies trying to model templates that use the general geometric properties of the masses [11], these are always complex, multiparameter models which are not applicable to all masses with various sizes and shapes.

In order to cope with the problems above, some researchers get inspirations from the doctors' detection process. They use the visual saliency to locate the suspicious lesions [14, 15]. Visual saliency models human beings' ability that perceives salient features in an image. In computer vision, these visual saliency-based methods compute probabilistic maps of an image where the pixels are very different from surrounding regions [16]. These methods greatly simplify the process of mass detection. For example, Tourassi et al. [14] proposed a novel similarity measure by incorporating the Gaussian salient map of image pixels. Agrawal et al. [16]

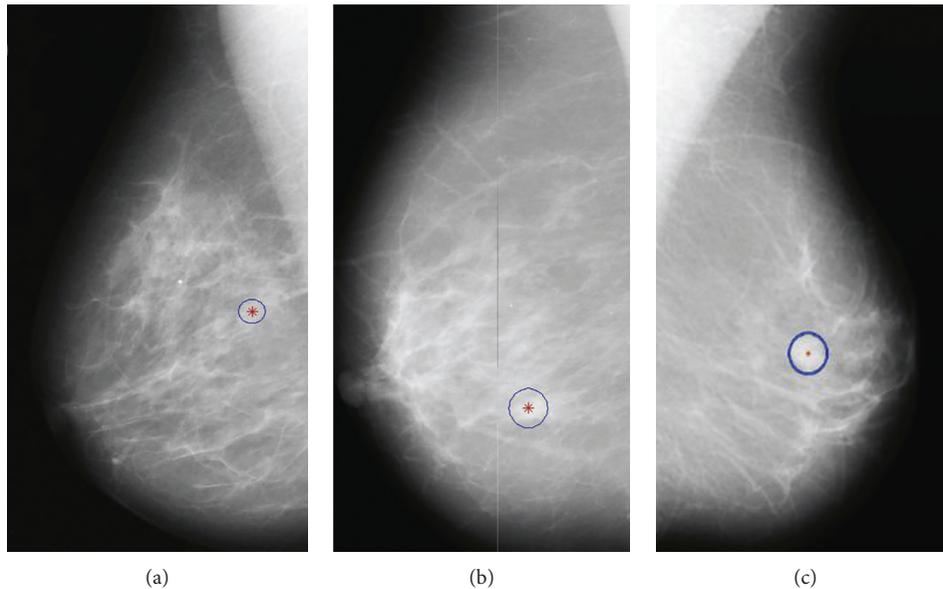


FIGURE 1: Sample images from MIAS data set [17].

proposed an automatic mass detection algorithm using the graph-based vision saliency (GBVS) map. However, most of these methods are derived from natural scene statistics while the characteristics of medical images are different. In fact, they only consider the size and brightness of suspicious regions, which is not enough for mass detection, as shown in Figure 1. Some masses are so small or they adhere tightly to glandular tissue. The visual significance is relatively low, resulting in a high false positive.

At present, deep learning [18] has been shown to be consistently producing higher performance compared with traditional machine learning methods [19]. Directly distilling information from training samples and convolutional neural networks (CNNs) [20] have been successfully applied to some medical tasks: for example, breast mass detection/diagnosis [21, 22], segmentation of the left ventricle [23], and classification of skin cancer [19]. Dhungel et al. [24] presented an automated mass detection method using a cascade of deep learning and random forest classifiers. Kooi et al. [25] have shown that a CNN trained on a large data set of around 45,000 images outperforms a state-of-the-art system in CAD. Most of these methods work well on large data sets but exhibit certain limitations on small data sets because they need to decide a large number of parameters [26]. Therefore, the traditional machine learning method is still valuable in some aspects, like data mining based on small samples [27], integration of multiple knowledge [28], and so on.

Studies have shown that the recognition of doctors plays an important role in lesion detection of radiology [29, 30]. Gestalt psychology tries to understand the laws of our ability to acquire and maintain meaningful perceptions in an apparently chaotic world [31]. Meanwhile, the theory has been validated on extensive experiments, which are performed on neatly organized dot lattices on a screen. These dots share many similarities with pixels in medical images [32, 33].

Hence, we consider incorporating visual perception properties described by the Gestalt psychology framework into mass detection. Inspired by Gestalt psychology, the Gestalt framework covers sensation (bottom-up) and perception (top-down), which are also collectively called recognition [34]. The theory aims to formulate visual rules according to which perceptual input is organized into unitary forms. The Gestalt theory includes the following principles: proximity, similarity, continuity, symmetry, closure, simplicity, and so on [34]. These visual rules can be used to help doctors to distinguish cancer masses from normal tissues.

Inspired by the framework of Gestalt theory, we propose to apply visual rules to medical image analysis. More exactly, we present an automatic mass detection framework based on Gestalt psychology. It contains three modules: sensation integration, semantic integration, and validation. In each module, a series of mathematical and calculation models for visual rules are presented. The proposed automatic mass detection method integrates human cognition properties and the visual characteristics of breast masses. To the best of our knowledge, combining bottom-up sensation and top-down recognition of the radiologist has not been attempted before. Experimental results demonstrated that the proposed method has yielded better performance than other algorithms.

## 2. Mass Detection Framework Inspired by Gestalt Psychology

In this paper, we propose to incorporate the visual perception properties into breast mass detection. First, the characteristics of mammogram reading by radiologists are analyzed as per Gestalt psychology. Second, a framework for automatic detection of masses is proposed. All visual rules in the framework are quantified through mathematical methods.

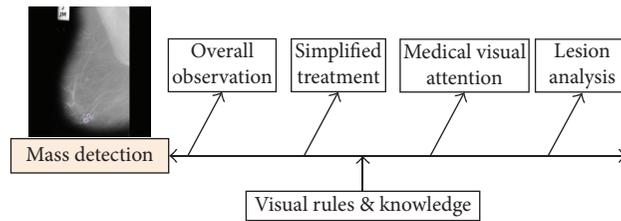


FIGURE 2: The clinical diagnosis of breast mass by the radiologist.

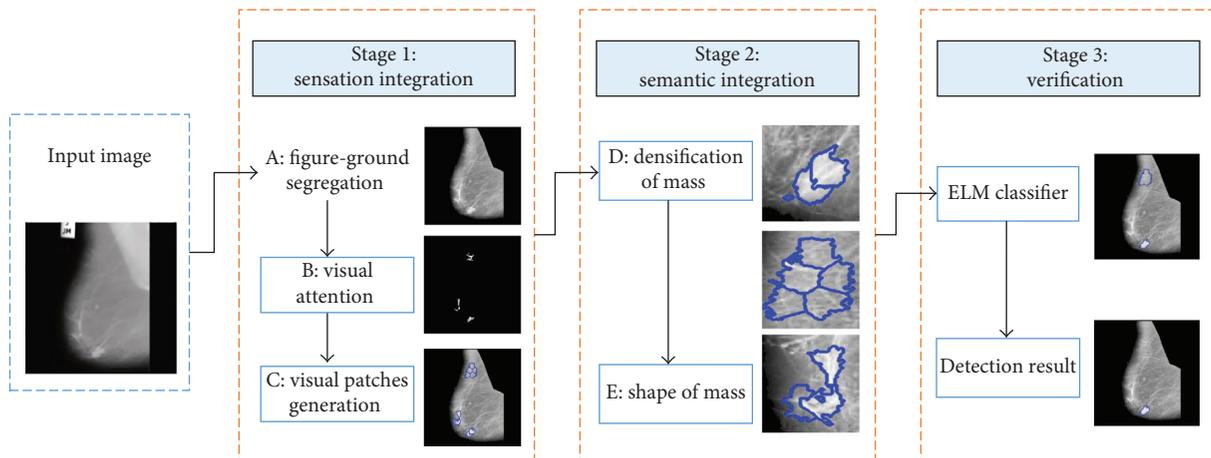


FIGURE 3: The framework of the proposed approach. It can be divided into three stages including sensation integration, semantic integration, and verification. Visual rules used in the framework are modeled and indicated with the labels A, B, C, D, and E.

**2.1. Analysis of Mammogram Screening under the Gestalt Framework.** In most cases, screening mammogram is a process of discovery, detection, and diagnosis by the radiologist [35]. The diversity of mammogram tissues brings many problems to mass detection. Radiologists are professional analysts who spend thousands of hours refining their abilities of detecting lesions in medical images. They gain a lot of experience in the learning process. According to Gestalt psychology research, radiologists read the images with the eyes and the brain. The visual rule plays an important role in recognizing masses from mammogram screening.

In clinical practice, radiologists tend to analyze medical images from overall impression down to individual parts for single-read mammography (or top-down), as shown in Figure 2. At the beginning, radiologists go through the mammogram and then focus on the highly suspicious areas. In vision psychology, the eyes can only accept a small number of associated units. If a Gestalt framework contains too many unrelated units, the eyes try to simplify it and combine the units into a big unit that can be processed easily. That is, our brain tends to combine and simplify the units [31]. Then, all the suspicious areas (called regions of interest, ROIs) will be further analyzed to get the masses. Many factors need to be considered by radiologists, such as morphology, density, and correlation with surrounding tissue. Generally speaking, a mass is a generic term indicating a localized swelling protuberance or lump in the breast [11].

**2.2. The Framework of Automatic Mass Detection Based on Gestalt Framework.** Inspired by the clinical practice, a mass detection method based on Gestalt framework is proposed in this paper (Figure 3). We divide the framework into three stages including sensation integration, semantic integration, and verification. It can be viewed as a combination of bottom-up sensation and top-down recognition methods. For each part, there are various visual rules based on Gestalt psychology and morphological characteristics.

**2.2.1. Stage 1: Sensation Integration.** In the initial stage, the visual sense of the radiologist plays an important role in mass detection. The attention is a process of selecting and getting visual information from pixels of the image (bottom-up) [36].

*Observation 1: From doctors' experience, the mass areas are located in the breast zone and are always more salient than the surrounding area. Inspired by the Gestalt framework, three rules are defined for image simplification, including figure-ground segregation, visual patches generation, and visual attention.*

**(1) Figure-Ground Segregation.** The principle of figure-ground segregation is one of the basic cognitive principles [37]. When applied to a mammogram, this principle supposes that the background does not contain any valuable information and is neglected by radiologists, obviously a simplified treatment. As a result, some existing methods propose to separate the figure from the background [38]. The

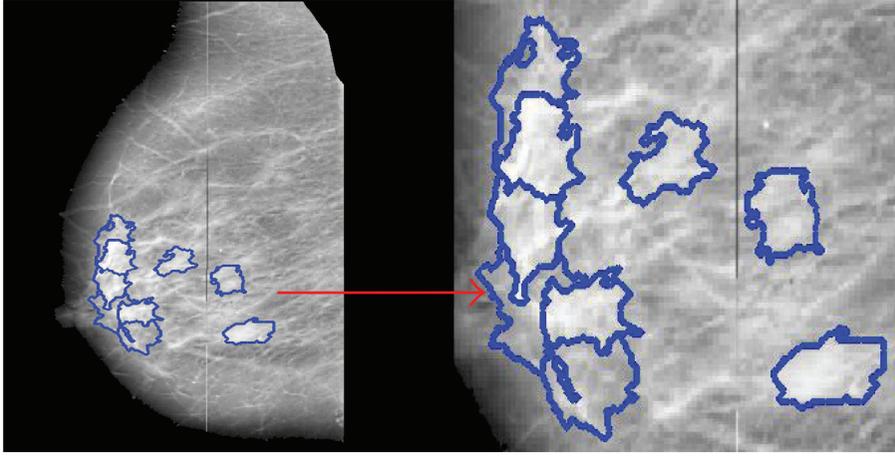


FIGURE 4: Visual patches based on Gestalt psychology.

mammograms have low contrast and still have noise in the background such as tape markings and labels (as shown in Figure 3). We use an adaptive global threshold to compute the outline of the breast region [39]. Based on the morphological analysis, an image enhancement method is adopted [40], which can effectively suppress the background and enhance the features of masses on mammograms simultaneously. Meanwhile, the pectoral muscle is removed [41].

(2) *Visual Attention in the Medical Image.* As we all know, the mass areas are always “brighter” than the surrounding areas. That is, the highlighted region attracts more visual attention than the background region when doctors browse the mammogram. It is called prominence in Gestalt psychology. A simple method is to predefine a threshold value for a breast image. However, this approach is unfavorable as there is a large variance between tissues in mammogram. Following [42], opening operation is adopted to find the focal area in the mammogram:

$$I_{\Phi} = I \circ \Phi = (I \otimes \Phi) \oplus \Phi, \quad (1)$$

where  $\circ$  is the morphological opening operation,  $\otimes$  is the erode operation, and  $\oplus$  is the dilation operation. The morphological opening operation consists of two steps in our method. Firstly, the original image  $I$  is eroded ( $\otimes$ ) with the structural element  $\Phi$ .  $\Phi$  is created by a flat disk-shaped structuring element with the specified radius of 6 pixels in the experiments. Secondly, dilation ( $\oplus$ ) is performed on the eroded image to produce the final reconstructed image ( $I_{\Phi}$ ). And then, we perform regional maximum on the reconstructed image, which retrieves all the salient regions without overselecting any of them.

(3) *Visual Patches Generation.* Gestalt theory aims to formulate some rules according to which the perceptual input is organized to unitary forms such as wholes, groups, or gestalt. The most common method is to group the similar or proximate object together. Inspired by the concept of “superpixel” [43], the basic processing units (visual patches) are generated by using our previous work [44]. In [44], the abdominal computed tomography (CT) image is divided into

many visual patches as per the law of similarity evaluated by both intensity and spatial distance. Now, the proposed method is applied on mammograms (Figure 4). Visual patches are generated by clustering pixels based on both intensity similarity and spatial proximity. Firstly,  $K$  cluster centers are set to divide the image into several rectangular patches. Then, we use the following similarity index  $D_s$  (2) to cluster pixels in mammograms:

$$D_s = \frac{\mu}{S} d_{xy} + d_g, \quad (2)$$

where  $d_{xy}$  is the spatial proximity,  $I_{\Phi} = I \circ \Phi = (I \otimes \Phi) \oplus \Phi$ , which is calculated by Euclidean distance on image plane,  $d_g$  is the intensity distance, and  $\mu$  is the parameter of the pixel compactness [44].  $S$  is a constant which balances the spatial proximity in image gray space, which is set as 25 in the experiment. The generated visual patches act as the basic processing units in the mass detection.

Visual patches are generated only in the salient positions which reduce the computation expense greatly as shown in Figure 4. These visual patches are the basic processing and analysis units in our algorithm. Suppose we have  $M$  visual patches which meet the condition of visual attention.  $U = \{P_1, P_2, \dots, P_M\}$  is the set of  $M$  visual patches in a mammogram, and we assume that all of these patches are candidate masses in the sensation phase. Meanwhile, we define another set  $N = \phi$  to store the patches which is regarded as normal tissue.

**2.2.2. Stage 2: Semantic Integration.** After the first stage (sensation integration), there are still many false-positive patches in  $U$  after the first stage (sensation integration). According to morphological characteristics of breast tissue, we introduce the semantic integration to further distinguish the masses patches from normal tissue. Kinds of semantic features of visual patches are used to help separate the mass from the normal tissue. The semantic integration can be regarded as a top-down recognition process [45].

*Observation 2: Normal tissue is always rich in glands and has poorer or lower density than that of the mass region. In addition,*

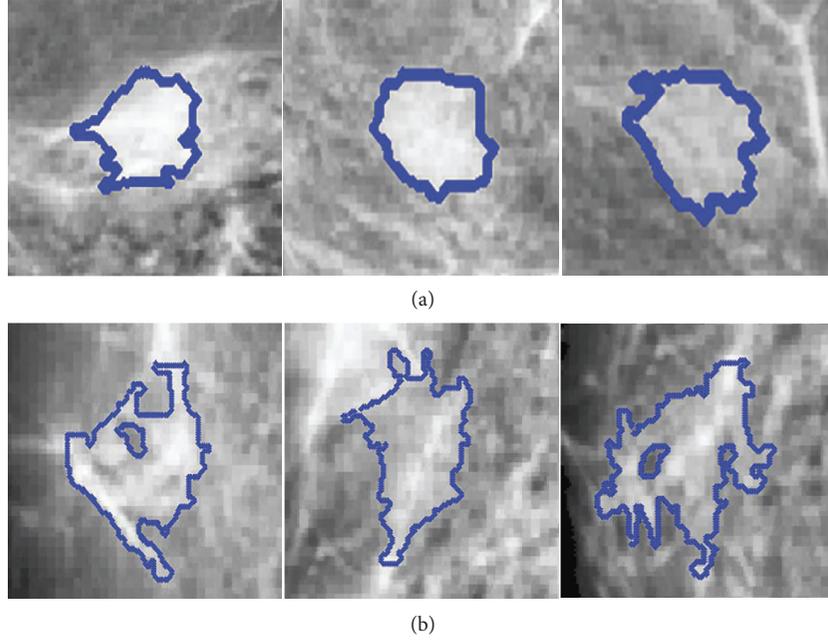


FIGURE 5: The densification of different patches. Visual patches of (a) mass and normal tissue.

the shape of masses tends to be round and oval [46, 47]. Masses and glandular tissues have very different shapes. Following the Gestalt principle, we propose two rules based on morphological characteristics of masses, including densification and shape.

(4) *Densification of Mass.* Eltonsy et al. [11] showed that the growth of a mass disrupted the normal breast parenchyma structure and formed a focal activity area called “focal seeds” in our research. According to the Gestalt principle, the law of similarity indicates that elements are perceptually grouped together if they are similar. The focal seed is regarded as an entirety because it has a strong self-similarity as shown in Figure 5(a), whereas normal tissue sometimes has a poor densification and even forms some holes in the visual patch (Figure 5(b)). The main reason is that normal tissue contains rich glands that affect the consistency of a given visual patch. The visual patches of masses exhibit homogeneity, whose solidity values are very high [48]. Here, the densification is defined as

$$\begin{aligned}
 \text{Dens} &= \sum_i \sum_j \frac{(ij)M(i, j) - u_i u_j}{\sigma_i \sigma_j}, \\
 u_i &= \sum_i \sum_j i \cdot M(i, j), \\
 u_j &= \sum_i \sum_j j \cdot M(i, j), \\
 \sigma_i^2 &= \sum_i \sum_j M(i, j)(i - u_i)^2, \\
 \sigma_j^2 &= \sum_i \sum_j M(i, j)(j - u_j)^2,
 \end{aligned} \tag{3}$$

where  $M(i, j)$  is the gray-level co-occurrence matrix of the visual patch  $P(i, j)$  and  $u$  and  $\sigma$  are the mean value and

variance for each visual patch  $P$ , respectively. The bigger the Dens (densification) is, the greater the probability of it being a tumor; otherwise, it is more likely to be the normal tissue. Therefore, a threshold  $T_{\text{dens}}$  is assigned based on the results of statistical analysis. Each visual patch in the candidate set  $U$  is detected, and the false-positive rate of visual patches is reduced as shown in (4), where  $N$  is the set to store visual patches of normal tissue:

$$\begin{aligned}
 P_i &\in N \text{ if } \text{Dens}(P_i) < T_{\text{dens}}, \\
 P_i &\in U \text{ if } \text{Dens}(P_i) \geq T_{\text{dens}},
 \end{aligned} \tag{4}$$

(5) *Shape of Mass.* As we all know, an abundance of glands exist in the breast. They are radically arranged from the center of nipple, like a crown, occupying a large part of the mammogram. The brightness of the gland is most similar to the tumor tissue, and this causes high false-positive results in various CAD systems [16]. But the shape of visual patches which contain glandular tissue is always a long strip as shown in Figure 6.

Comparatively, the visual patches of tumor tend to be round and oval [46] (Figure 5(a)). In addition, the focused visual patch is positioned only in the center area of mass rather than its edge. The main reason is that many masses have no obvious margins surrounded by glandular tissues. The continuity law of Gestalt states that elements of objects tend to be grouped together and integrated into perceptual wholes. In other words, the patches containing glandular tissue can be easily perceived with the distribution of continuity. In that case, the linear patches can be filtered from the candidate visual patches.

Here, we bring in the concept of eccentricity to restrict the shape of each visual patch, which is an important index in ellipse. An ellipse fitting algorithm is used for each visual

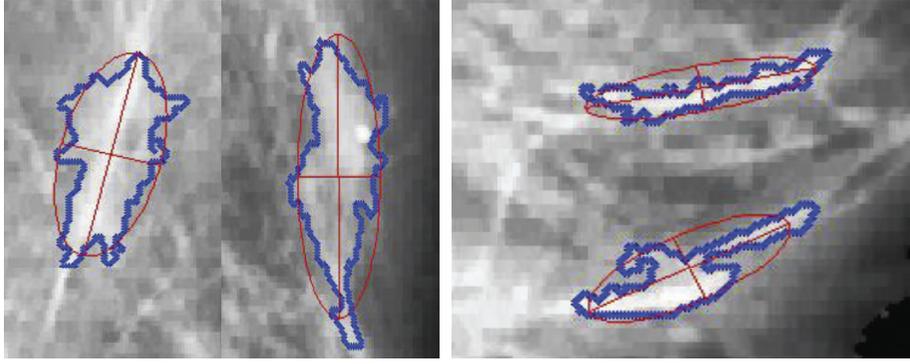


FIGURE 6: Visual patches with the glandular.

patch edge to get the eccentricity (Figure 6). The elliptic equation is defined as

$$A'x^2 + B'xy + C'y^2 + D'x + E'y + F' = 0. \quad (5)$$

There are 6 parameters in (5), that is,  $A'$ ,  $B'$ ,  $C'$ ,  $D'$ ,  $E'$ , and  $F'$ . They can be estimated by using the least-squares method according to the edge points of each patch. Then, the eccentricity is defined using the long and minor axis of the fitted ellipse:

$$E(P_i) = \frac{\sqrt{a_i^2 - b_i^2}}{a_i} = \sqrt{1 - \left(\frac{b_i}{a_i}\right)^2},$$

$$a_i = 2\sqrt{\frac{-2F'}{A' + C' - \sqrt{B'^2 + (A' - C'/F')^2}}},$$

$$b_i = 2\sqrt{\frac{-2F'}{A' + C' + \sqrt{B'^2 + (A' - C'/F')^2}}}, \quad (6)$$

where  $a_i$  and  $b_i$  are margin axis and minor axis of the ellipse, respectively.  $E(P_i)$  measures the circularity of the visual patch  $P_i$  and the range is from 0 to 1. Since we assume that the eccentricity value of a visual patch is significantly lower in tumors than that in normal tissues. A threshold  $T_e$  is defined to reduce the false-positive rate:

$$\begin{aligned} P_i &\in N \text{ if } E(P_i) \geq T_e, \\ P_i &\in U \text{ if } E(P_i) < T_e. \end{aligned} \quad (7)$$

Based on the above rules, the number of candidate patches in set  $U$  is greatly reduced, while the sensitivity and specificity are very high which will be explained in detail in the experiment. Another advantage of our method is that the focused visual patches only lie in the center of tumors. It can lay the foundation for the further proceeding and analysis, such as mass segmentation and mass diagnosis.

### 2.2.3. Stage 3: Verification: False-Positive Reduction Based on Texture Feature Classification

Observation 3: *The Gestalt theory shows that the information*

*of visual perception is affected by observer's prior experience. Radiologists accumulate a great deal of knowledge to identify the texture of breast tissue. We propose a validation method based on texture of mass and extreme learning machine (ELM) classification method.*

The importance of texture features is obvious for mass detection. In this research, we have extracted intensity and gray-level co-occurrence matrix (GLCM) features on the candidate visual patches, including  $N$  dimensions for the patch  $P_i$ ,  $F(P_i) = \{g_1, g_2, \dots, g_N\}$ . Among these features, two of them are the first-order texture feature, which is the mean of gray value and gray variance describing visual patches. Ten features are calculated from the gray-level co-occurrence matrix, namely, contrast, correlation, energy, homogeneity, standard deviation, inverse difference movement, kurtosis, skewness, entropy, and root mean square.

The candidate patches left in  $U$  serve as the seeds of region growing algorithm. ROIs are obtained by clustering these visual patches based on texture features  $F$ . The features of each ROI are represented by the average value of associated patches as shown in (8), where  $M_j$  is the total number of visual patches in ROI $_j$ :

$$F(\text{ROI}_j) = \frac{\sum_{P_i \in \text{ROI}_j} F(P_i)}{M_j}. \quad (8)$$

Recently, extreme learning machine (ELM) has been extensively studied, and important progress has been made in both theories and practical applications. The ELM theory in [49] has proved that random feature mapping can provide universal approximation capability. The ELM has built some tangible links between machine learning techniques and biological learning mechanisms. It is an emerging learning algorithm for the generalized hidden layer feedforward neural network [49, 50]. Here, the ELM is used to simulate the final decision of doctors. Furthermore, it classifies the ROIs into normal and abnormal cases based on the texture feature extraction.

## 3. Experimental Results

**3.1. Data Set and Parameter Setting.** Our proposed method is tested on two publicly available data sets: MIAS [17] and DDSM [51]. A set of 257 mammograms of MIAS is used in the research. Among these images, 207 images do not

contain any lesions while other 50 images have masses. The spatial resolution of image in MIAS is  $50 \mu\text{m} \times 50 \mu\text{m}$ , and grayscale intensity is quantized to 8 bits. The DDSM data set contains 210 images, in which 130 images contain masses and the other ones are normal mammograms. The images of DDSM have been resized to  $1024 \times 1024$  pixels, and grayscale intensity is quantized to 8 bits in accordance with images in MIAS. In both MIAS and DDSM data sets, the mammograms containing masses have been annotated by expert radiologists, which are used for reporting the detection performance in our experiments. The ELM classification method divides visual patches into mass and nonmass candidates using 10-fold cross validation.

In our research, a series of indexes are used to quantitatively evaluate the effectiveness of our method. The performance indexes include sensitivity (Sens), false positives per image (FPI), homogeneity (Dens), and free-response receiver operating characteristic (FROC). These indexes are described below.

- (a) Sens and FPI are computed as a region-based evaluation. If the ratio of the overlapping region of the ground truth and the computer-segmented mass region is larger than 50%, the region is considered as “True Positive Marks” or “Positive ROIs” [52]. Otherwise, it is considered as “False Positive Marks,” which is also called “Negative ROIs” in our experiments. For computer-aided systems, we would like the value of sensitivity to be as high as possible. Meanwhile, FPI should be low while guaranteeing high sensitivity [53].

$$\text{Sens} = \frac{\text{number of true positive marks}}{\text{number of regions}} \quad (9)$$

$$\text{FPI} = \frac{\text{number of false positive marks}}{\text{number of images}}.$$

- (b) In Section 3.2, we use homogeneity (Dens) to characterize the distribution of visual patches, which is defined as (3). It is based on the fact that every visual patch is an independent processing unit that should be homogeneous as per Gestalt rules [44]. The value of Dens ranges from 0 to 1. When patches are uniform, the value of Dens tends to be 1, while for nonuniform patches, the value tends to be 0.
- (c) In Section 3.5, free-response receiver operating characteristic (FROC) [52] curve is used. The FROC curve is defined as the plot of sensitivity (Sens) versus the average number of false positives per image (FPI).

All numerical methods are performed using MATLAB 2012b software running on a desktop PC with a 2.50 GHZ CPU and 2G RAM. Different from data-driven algorithms like deep learning, our method does not need a large amount of data. The major reason is that it is designed based on human visual characteristic and radiologists’ experience. There are all together three types of parameters in our

TABLE 1: Three types of parameters in the experiments.

Type	Parameters	Setting value
Visual rules	Medical visual attention threshold range: [0 1]	0.5
	Densification of the mass threshold range: [0 1]	0.93
	Shape of mass threshold range: [0 1]	0.86
Generation of visual patch	Image size	[1024, 1024]
	Cluster parameter, $K$	2000
ELM	Tradeoff parameters	$C = 10$
	Kernel type	RBF
	Kernel parameter	0.01

method, which are (descending order of importance) the thresholds for visual rules, the parameters for generating visual patch, and some other parameters of the ELM classifier.

The medical images from different hospitals or different apparatuses may be completely different. Thus, the parameters of the method should be modified on different apparatuses or data sets. The thresholds for visual rules (medical image attention, densification of mass, and shape of mass) are crucial in the proposed method, which determine the number of suspicious regions, because three key parameters in the method are independent. By comparing the Sens and FPI in different thresholds, three parameters can be determined for new validation data. In addition, the parameters for visual patches generation and ELM classifier have less impact on detection performance than that for visual rules. If the input size of image is  $[M, N]$ , then the initial size of the visual patch can be calculated by  $M * N/K$ , where  $K$  is the cluster parameter for visual patches generation. That is to say, the bigger the value of  $K$  is, the smaller the size of visual patches is, and vice versa. The initial size of the visual patch is preferably less than the size of the entire breast mass. In this manuscript, both  $M$  and  $N$  is 1024, and  $K$  is set as 2000. ELM has been extensively studied, and it shows a good convergence speed and stability [49]. In our research, the parameters of ELM do not require careful adjustment. All parameters used in the experiments are shown in Table 1.

*3.2. The Effectiveness of Visual Patches.* The mass region patches on mammogram are called “positive patches,” and the normal tissue ones are “negative patches” in the following content. Figure 7 shows the statistical histogram of homogeneity of visual patches both for negative and positive patches. It is obvious that the homogeneity of all patches is above 0.85, with positive patches having a homogeneity value above 0.9, center around the value 0.95. We can draw a conclusion that the use of visual patches ensures the semantic consistency of objects in ROIs, which lays a solid foundation for further research. Besides, the homogeneity of positive patches is a bit higher than that of negative patches, which supports the description of similarity rules defined in Section 2.2.2. As shown in Figure 7, the homogeneity distribution curves of positive and negative patches are similar

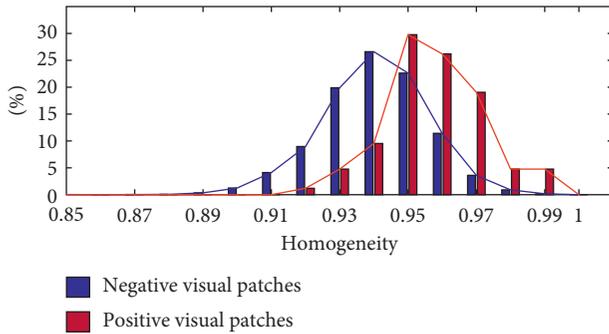


FIGURE 7: The statistical histogram of homogeneity of negative and positive visual patches.

to normal distribution. The two curves are distinguishable owing to the higher homogeneity of positive patches than that of the negative ones.

**3.3. Mass Detection Performance.** In the proposed framework, a series of visual rules have been defined, and finally, all the salient patches are saved (in set  $U$ ). We compare the performance of our framework to three existing visual saliency algorithms: Agrawal et al. [16] (graph-based vision saliency), Achanta and Ssstrunk [54] (maximum symmetric surround saliency), and Murray et al. [55] (saliency estimation using a nonparametric low-level vision model), as shown in Figure 8.

In Figure 8, each row corresponds to the output of four algorithms, and the corresponding mammograms are shown in the first column. The detection results of three stages in the proposed method have been listed in Figure 8(d). It is obvious that our method outperforms other saliency algorithms for mammogram mass detection. At the stage of verification, the false-positive rate is further reduced and there is a bit impact on the mass region compared with normal tissue. We experimentally observe that Agrawal et al. [16] yielded a relatively good result that is in accordance with what is reported in [16]. However, it computes the saliency of a region with respect to its local neighborhood using the directional contrast. But it is invalid when the mass is surrounded by dense gland tissues as shown in the last row of Figure 8. However, the positive aspect of our method is that it combines visual cognitive theory with various morphological characteristics of masses. The advantages can be summarized as follows: (1) The detection method based on gestalt rules is able to detect masses of varying sizes, resulting in a low false-positive rate (the green region in Figure 8). (2) The salient visual patches of our method mostly lie in center of the ground truth regions. The detected results based on the proposed method can be used for further analysis, such as mass segmentation and diagnosis.

Table 2 shows the overall performance of our proposed method on MIAS and DDSM data set. As shown in Table 2, the performances of three stages of our method are given. Masses are detected effectively, and Sens reaches 92% at 1.94 FPI and 93.84% at 2.21 on MIAS and DDSM data sets, respectively. As shown in Table 3, we can get extended statistic information on both the total number and

percentage of patches and ROIs. Meanwhile, the performance curves are plotted in Figure 9. From Table 3 and Figure 9, the number of positive patches and positive ROIs remains largely unchanged, whereas the number of negative examples is greatly reduced as the detection stages are introduced. The performance curves show the similar change of detection performance for MIAS and DDSM. We can draw a conclusion that the positive visual patches can be preserved, and false-positive results are declined dramatically under the gestalt rules constraint.

**3.4. Influence of the Breast Density for Mass Detection Performance.** In general, masses in low-density breast are easily detected than masses in high-density breast [56]. To further evaluate the ability of our method, mass images with different densities are separated to test on the MIAS and DDSM data sets. The results are summarized in Table 4. There are three types of densities for MIAS, that is, fatty ( $F$ ), fatty-glandular ( $G$ ), and dense-glandular ( $D$ ). Different from MIAS, the images in DDSM data set are divided into four categories based on breast density, that is, 1, 2, 3, and 4.

The quantity proportions of each subset with different densities are listed in the first row of Table 4. This table shows Sens and FPI of mass images with different densities. It shows that the algorithm has good performance and works well on different mass images. Looking at the MIAS results, the proposed method has superior performance on fatty ( $F$ ) and fatty-glandular ( $G$ ) breast images compared to the dense-glandular ( $D$ ) images. Similarly, the method performs better in low-density (level 1 and level 2) images than high-density (level 3 and level 4) images on DDSM. Masses in low-density images usually have distinctive visual features compared with the tissue around it. Conversely, some visual patches with high density may cause erroneous judgment at the final recognition stage. So, the false-positive rates would increase when the data set includes many breast images with high density. In this section, the sensitivities for all lesions are 92% at 1.94 FPI on MIAS and 93.84% at 2.21 FPI on DDSM.

A common method for evaluating true-positive detection is free-response receiver operating characteristic (FROC) analysis [57]. It is a plot of operating points showing a tradeoff between the sensitivity rate and the average number of false positives per image. The complete FROC curves of our method are presented in Figure 10. The blue and red curves denote the detection results on MIAS and DDSM, respectively. We can get a favorable detection result when a false positive reaches 2 per image on the two data sets.

**3.5. Comparison of Experimental Results.** The mass saliency is introduced in the proposed framework. In order to evaluate the ability of our method, experiments were conducted with 10-fold cross validation on MIAS and DDSM. The performance is compared with other popular algorithms in terms of Sens and FPI. It is clearly shown in Table 5 that the proposed method has equivalent or even better accuracy

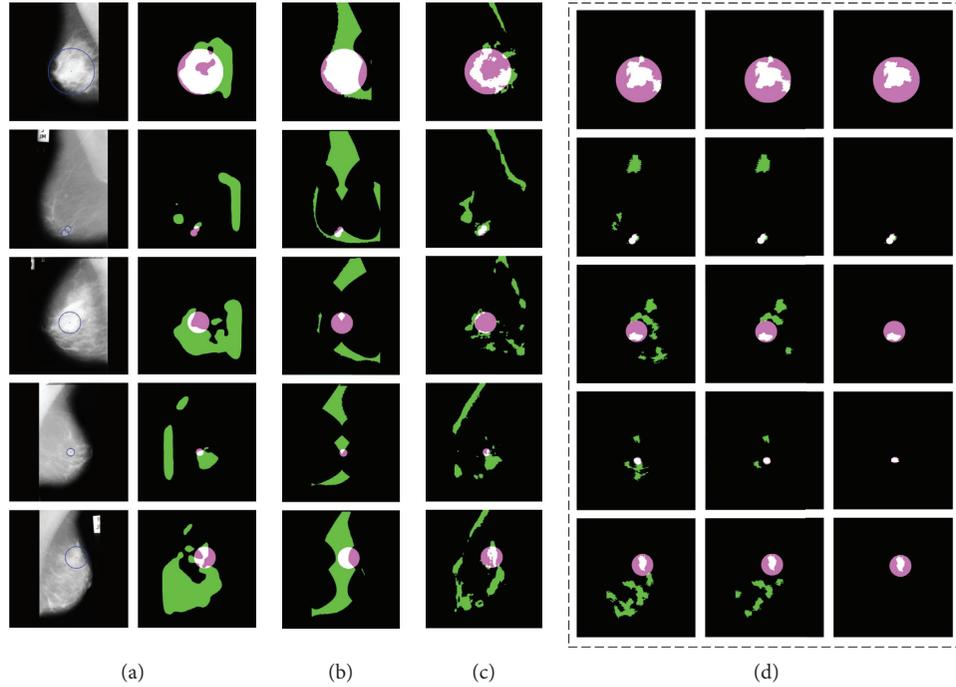


FIGURE 8: Sample results of the saliency algorithms. Green denotes the saliency region detected by these algorithms, pink represents the ground truth region containing mass, and white denotes the crossing region between green and pink. (a) Agrawal et al. [16], (b) Achanta and Süsstrunk [55], (c) Murray et al. [56], and (d) the three stages of our method. Stage 1: the fifth column is the detection result of sensation integration. Stage 2: the sixth column is the detection result of semantic integration. Stage 3: the last column is the final detection result (verification) of our method.

TABLE 2: Mass detection performance (Sens and FPI) of different stages in our method.

Data	Index	Stage 1		Stage 2		Stage 3
		Visual attention		Densification	Shape	Texture
MIAS	Sens	100%	98%	96%	92%	
	FPI	4.12	4.21	2.82	1.94	
DDSM	Sens	99.23%	96.92%	96.15%	93.84%	
	FPI	4.78	4.19	3.01	2.21	

TABLE 3: Mass detection performance (number and percentage of patches/ROIs) of different stages in the proposed method.

Data	Unit	Stage 1		Stage 2		Stage 3			
		Visual attention		Densification	Shape	Texture			
MIAS	Total patches	864	100%	684	79.10%	518	59.95%	218	25.23%
	Positive patches	82	100%	77	93.9%	75	91.46%	71	86.59%
	Negative patches	782	100%	607	77.62%	443	56.67%	147	18.80%
	Total ROIs	256	100%	248	96.88%	189	73.83%	143	55.86%
	Positive ROIs	50	100%	50	100%	48	96%	46	92%
	Negative ROIs	206	100%	198	96.12%	141	68.45%	97	47.09%
DDSM	Total patches	2300	100%	1397	60.74%	1074	46.70%	671	29.17%
	Positive patches	198	100%	185	93.43%	180	90.91%	161	81.31%
	Negative patches	2102	100%	1212	57.66%	894	42.53%	510	24.26%
	Total ROIs	766	100%	690	90.08%	528	68.93%	418	54.57%
	Positive ROIs	144	100%	142	98.61%	137	95.14%	131	90.97%
	Negative ROIs	622	100%	548	88.10%	391	62.86%	287	46.14%

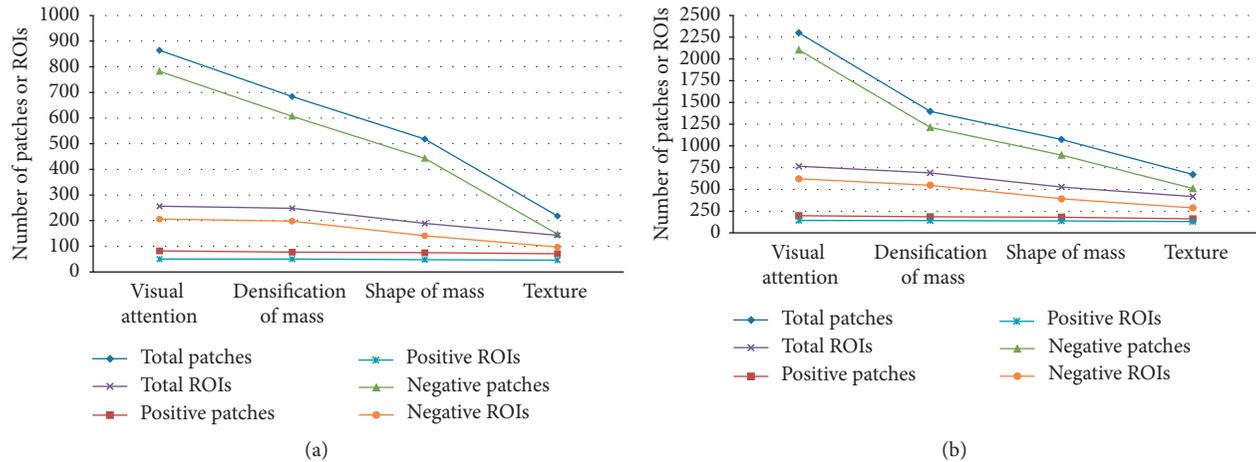


FIGURE 9: The number and percentage of patches/ROIs are counted for each step of our method: (a) plotted on the MIAS data set and (b) plotted on the DDSM data set.

TABLE 4: Influence of the breast density on the proposed detection algorithm.

MIAS					
Density	<i>F</i>	<i>G</i>	<i>D</i>	—	All
Proportion	44%	36%	20%	—	100%
Sens	90.9%	94.44%	90%	—	92%
FPI	1.77	2.05	2.1	—	1.94
DDSM					
Density	1	2	3	4	All
Proportion	7.69%	23.84%	36.92%	31.53%	100%
Sens	90%	96.77%	93.75%	92.68%	93.84%
FPI	1.2	2.0	2.08	2.75	2.21

The first row shows the quantity proportion of each subset with different densities on both MIAS and DDSM data sets. The metrics are Sens and FPI.

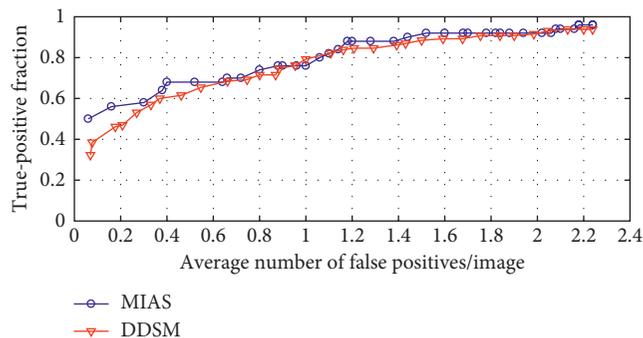


FIGURE 10: FROC curves of the proposed method on MIAS and DDSM data sets.

than other algorithms. High sensibility and low FPI represent the good performance of an algorithm. We can get a series of Sens at different FPIs as shown in Figure 10. The sensibility reaches 92% at 1.94 FPI or 94% at 2.16 FPI on MIAS. Accordingly, the sensibility reaches 93.84% at 2.21 FPI or 94.6% at 2.66 FPI on DDSM. The comparative studying methods [52, 53, 58], following the similar pipeline, include two parts: image preprocessing and suspicious mass

TABLE 5: Comparing the performance (Sens and FPI) of the proposed method with existing algorithms on the MIAS and DDSM data sets.

Algorithm	Data set	Sens	FPI
Wavelet processing and adaptive threshold [51]	MIAS	90.9%	2.35
Havrda and Charvat entropy and OSTU [52]	MIAS	93.2%	7.6
Adaptive median filtering and texture analysis [58]	MIAS	92.3%	2.75
Our method (Gestalt psychology)	MIAS	94%	2.16
Wavelet processing and adaptive threshold [51]	DDSM	91%	2.1
Kernelized supervised hashing [59]	DDSM	94%	4.1
Our method (Gestalt psychology)	DDSM	94.6%	2.66

regions identification with different adaptive thresholds. In contrast, a sliding window scheme is utilized in [59], and texture features are modeled by kernel-based supervised hashing to get the mass location. Different from the sliding window in [59], visual attention of radiologists is used in our method. In the end, our method not only utilizes the advantages of machine learning approaches, but the visual saliency of mass is also modeled which achieves significant improvement in reducing false positives and sensitivity.

## 4. Conclusion

In this paper, we have proposed an automatic mass detection framework for digitized mammograms. The main contributions of our research can be summarized as follows: (1) The visual characteristic of radiologists is modeled based on the Gestalt theory. (2) An automatic mass detection framework is proposed which is in accordance with the doctors' visual perception. Some constraints are defined such as density and shape of visual patches. These parameters are probably used by experienced radiologists in detection/diagnosis masses and so on. (3) Our framework achieves a good performance both on MIAS and DDSM data sets.

Different from existing detection techniques, our methods use the visual patches as the basic processing unit. We focus on providing a more efficient and more innovative data analysis method for lesions detection than traditional methods. In our future work, more mammograms from different apparatuses will be tested to evaluate the performance of our proposal. We will further improve our method as per the feedback reports from more radiologists. Moreover, we would like to focus on expanding this research to the deep learning method.

## Conflicts of Interest

The authors would like to confirm that no conflicts of interest, financial or others, exist.

## Acknowledgments

This work was supported by the Major Program of National Natural Science Foundation of China (no. 81727802), the National Natural Science Foundation of China (nos. 61701404 and 81671648), the key problems of Social Development Science and Technology of Shaanxi Province (no. 2015SF119), the funds of clinical major projects of Tangdu Hospital (no. 2015LCYJ001), and the Graduate Students Innovative Talents Training Program of Northwest University (no. YZZ15095).

## References

- [1] A. Jemal, R. Siegel, E. Ward et al., "Cancer statistics, 2008," *CA: A Cancer Journal for Clinicians*, vol. 58, no. 2, pp. 71–96, 2008.
- [2] Z. Wang, G. Yu, Y. Kang, Y. Zhao, and Q. Qu, "Breast tumor detection in digital mammography based on extreme learning machine," *Neurocomputing*, vol. 128, no. 5, pp. 175–184, 2014.
- [3] A. U. Rehman, N. Chouhan, and A. Khan, "Diverse and discriminative features based breast cancer detection using digital mammography," in *Proceedings of the 2015 13th International Conference on Frontiers of Information Technology (FIT)*, pp. 234–239, Islamabad Pakistan, December 2015.
- [4] N. R. Mudigonda, R. M. Rangayyan, and J. L. Desautels, "Detection of breast masses in mammograms by density slicing and texture flow-field analysis," *IEEE Transactions on Medical Imaging*, vol. 20, no. 12, pp. 1215–1227, 2001.
- [5] J. Chu, H. Min, L. Liu, and W. Lu, "A novel computer aided breast mass detection scheme based on morphological enhancement and SLIC superpixel segmentation," *Medical Physics*, vol. 42, no. 7, pp. 3859–3869, 2015.
- [6] A. Mencattini and M. Salmeri, "Breast masses detection using phase portrait analysis and fuzzy inference systems," *International Journal of Computer Assisted Radiology and Surgery*, vol. 7, no. 4, pp. 573–583, 2012.
- [7] N. Dhungel, G. Carneiro, and A. P. Bradley, "Fully automated classification of mammograms using deep residual neural networks," in *Proceedings of the IEEE International Symposium on Biomedical Imaging*, pp. 310–314, Melbourne, Australia, April 2017.
- [8] C. Gallego-Ortiz and A. L. Martel, "Improving the accuracy of computer-aided diagnosis for breast MR imaging by differentiating between mass and nonmass lesions," *Radiology*, vol. 278, no. 3, pp. 679–688, 2015.
- [9] N. Karssemeijer and G. M. te Brake, "Detection of stellate distortions in mammograms," *IEEE Transactions on Medical Imaging*, vol. 15, no. 5, pp. 611–619, 1996.
- [10] G. Kom, A. Tiedeu, and M. Kom, "Automated detection of masses in mammograms by local adaptive thresholding," *Computers in Biology and Medicine*, vol. 37, no. 1, pp. 37–48, 2007.
- [11] N. H. Eltonsy, G. D. Tourassi, and A. S. Elmaghraby, "A concentric morphology model for the detection of masses in mammography," *IEEE Transactions on Medical Imaging*, vol. 26, no. 6, pp. 880–889, 2007.
- [12] P. Casti, A. Mencattini, M. Salmeri et al., "Contour-independent detection and classification of mammographic lesions," *Biomedical Signal Processing and Control*, vol. 25, pp. 165–177, 2016.
- [13] N. Petrick, H.-P. Chan, B. Sahiner, and D. Wei, "An adaptive density-weighted contrast enhancement filter for mammographic breast mass detection," *IEEE Transactions on Medical Imaging*, vol. 15, no. 1, pp. 59–67, 1996.
- [14] G. D. Tourassi, B. Harrawood, S. Singh, J. Y. Lo, and C. E. Floyd, "Evaluation of information-theoretic similarity measures for content-based retrieval and detection of masses in mammograms," *Medical Physics*, vol. 34, no. 1, pp. 140–150, 2007.
- [15] W. E. Polakowski, D. A. Cournoyer, S. K. Rogers et al., "Computer-aided breast cancer detection and diagnosis of masses using difference of gaussians and derivative-based feature saliency," *IEEE Transactions on Medical Imaging*, vol. 16, no. 6, pp. 811–819, 1997.
- [16] P. Agrawal, M. Vatsa, and R. Singh, "Saliency based mass detection from screening mammograms," *Signal Processing*, vol. 99, no. 6, pp. 29–47, 2014.
- [17] J. Suckling, J. Parker, D. Dance et al., "The mammographic image analysis society digital mammogram database," *International Congress Series Exerpta Medica*, vol. 1069, pp. 375–378, 1994.
- [18] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [19] A. Esteva, B. Kuprel, R. A. Novoa et al., "Corrigendum: dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.
- [20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *International Conference on Neural Information Processing Systems*, pp. 1097–1105, Doha, Qatar, November 2012.
- [21] N. Dhungel, G. Carneiro, and A. P. Bradley, "A deep learning approach for the analysis of masses in mammograms with minimal user intervention," *Medical Image Analysis*, vol. 37, pp. 114–128, 2017.
- [22] J. Arevalo, M. A. G. Lopez, and M. A. G. Lopez, "Representation learning for mammography mass lesion classification with convolutional neural networks," *Computer Methods and Programs in Biomedicine*, vol. 127, pp. 248–257, 2016.
- [23] T. A. Ngo, Z. Lu, and G. Carneiro, "Combining deep learning and level set for the automated segmentation of the left ventricle of the heart from cardiac cine magnetic resonance," *Medical Image Analysis*, vol. 35, p. 159, 2017.
- [24] N. Dhungel, G. Carneiro, and A. P. Bradley, "Automated mass detection in mammograms using cascaded deep learning and random forests," in *International Conference on Digital Image Computing Techniques and Applications*, pp. 1–8, Gold Coast, Australia, November–December 2016.
- [25] T. Kooi, G. Litjens, G. B. Van et al., "Large scale deep learning for computer aided detection of mammographic lesions," *Medical Image Analysis*, vol. 35, pp. 303–312, 2016.

- [26] G. Litjens, T. Kooi, B. E. Bejnordi et al., "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, no. 9, pp. 60–68, 2017.
- [27] G. Xu, Y. Ding, C. Wu, Y. Zhai, and J. Zhao, "Explore maximal frequent item sets for big data pre-processing based on small sample in cloud computing," in *International Congress on Ultra Modern Telecommunications and Control Systems and Workshops*, pp. 235–239, Brno, Czech Republic, October 2016.
- [28] K. C. Lee, J. H. Han, Y. U. Song, and W. J. Lee, "A fuzzy logic driven multiple knowledge integration framework for improving the performance of expert systems," *Intelligent Systems in Accounting Finance and Management*, vol. 7, no. 4, pp. 213–222, 2015.
- [29] T. Drew, K. Evans, M. L.-H. Vö, F. L. Jacobson, and J. M. Wolfe, "Informatics in radiology: what can you see in a single glance and how might this guide visual search in medical images?," *Radiographics*, vol. 33, no. 1, pp. 263–274, 2013.
- [30] W. Huda and K. M. Ogden, "How do radiographic techniques affect mass lesion detection performance in digital mammography?," *Proceedings of the SPIE*, vol. 5372, pp. 372–382, 2004.
- [31] F. Jäkel, M. Singh, F. A. Wichmann, and M. H. Herzog, "An overview of quantitative approaches in gestalt perception," *Vision Research*, vol. 126, pp. 3–8, 2016.
- [32] Y. Kinoshita, M. Koppen, and K. Yoshida, "Perception of image similarity considering gestalt theory," in *Proceedings of the International Conference of Soft Computing and Human Sciences*, pp. 171–177, Kitakyushu, Japan, August 2007.
- [33] S. H. Kang, "From gestalt theory to image analysis: a probabilistic approach [book review of mr2365342]," *Siam Review*, vol. 52, no. 2, pp. 399–400, 2010.
- [34] J. Wagemans, J. Feldman, S. Gepshtein et al., "A century of gestalt psychology in visual perception: II. Conceptual and theoretical foundations," *Psychological Bulletin*, vol. 138, no. 6, pp. 1172–1217, 2012.
- [35] H. Weedon-Fekjær, P. R. Romundstad, and L. J. Vatten, "Modern mammography screening and breast cancer mortality: population study," *British Medical Journal*, vol. 348, pp. g3701–g3708, 2014.
- [36] U. Rutishauser, D. Walther, C. Koch, and P. Perona, "Is bottom-up attention useful for object recognition?," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 37–44, Washington, DC, USA, July 2004.
- [37] S. L. Fischer, "The gestalt research tradition: figure and ground," *Gestalt Review*, vol. 16, no. 1, pp. 3–6, 2012.
- [38] C.-C. Jen and S.-S. Yu, "Automatic detection of abnormal mammograms in mammographic images," *Expert Systems with Applications*, vol. 42, no. 6, pp. 3048–3055, 2015.
- [39] P. Casti, A. Mencattini, M. Salmeri et al., "Estimation of the breast skin-line in mammograms using multidirectional Gabor filters," *Computers in Biology and Medicine*, vol. 43, no. 11, pp. 1870–1881, 2013.
- [40] D. S. Gowri and T. Amudha, "A review on mammogram image enhancement techniques for breast cancer detection," in *International Conference on Intelligent Computing Applications (ICICA)*, pp. 47–51, Pune, India, December 2014.
- [41] R. J. Ferrari, R. M. Rangayyan, J. L. Desautels, R. Borges, and A. F. Frere, "Automatic identification of the pectoral muscle in mammograms," *IEEE Transactions on Medical Imaging*, vol. 23, no. 2, pp. 232–245, 2004.
- [42] R. Bharath, L. Z. J. Nicholas, and X. Cheng, "Scalable scene understanding using saliency-guided object localization," in *10th IEEE International Conference on Control and Automation (ICCA)*, pp. 1503–1508, Hangzhou, China, June 2013.
- [43] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.
- [44] P. Li, J. Feng, Q. Bu, F. Liu, and H. Wang, "Multi-object segmentation for abdominal CT image based on visual patch classification," in *CCF Chinese Conference on Computer Vision*, pp. 130–138, Springer, Berlin, Germany, 2015.
- [45] X. He, R. S. Zemel, and D. Ray, "Learning and incorporating top-down cues in image segmentation," in *Lecture Notes in Computer Science*, vol. 3951, pp. 338–351, Springer, Berlin, Germany, 2006.
- [46] I. Vizcano, L. Gadea, L. Andreo et al., "Short-term follow-up results in 795 nonpalpable probably benign lesions detected at screening mammography," *Radiology*, vol. 219, no. 2, pp. 475–483, 2001.
- [47] R. Salvador, J. L. Lirola, R. Domnguez, M. López, and N. Risueño, "Pseudo-angiomatic stromal hyperplasia presenting as a breast mass: imaging findings in three patients," *Breast*, vol. 13, no. 5, pp. 431–435, 2004.
- [48] X. Gao, Y. Wang, X. Li, and D. Tao, "On combining morphological component analysis and concentric morphology model for mammographic mass detection," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 2, pp. 266–273, 2010.
- [49] J. Tang, C. Deng, and G.-B. Huang, "Extreme learning machine for multilayer perceptron," *IEEE transactions on neural networks and learning systems*, vol. 27, no. 4, pp. 809–821, 2016.
- [50] G.-B. Huang, "An insight into extreme learning machines: random neurons, random features and kernels," *Cognitive Computation*, vol. 6, no. 3, pp. 376–390, 2014.
- [51] M. Heath, K. Bowyer, D. Kopans, R. Moore, and W. P. Kegelmeyer, "The digital database for screening mammography," in *Proceedings of the 5th International Workshop on Digital Mammography*, pp. 212–218, Toronto, Canada, June 2000.
- [52] P. S. Vikhe and V. R. Thool, "Mass detection in mammographic images using wavelet processing and adaptive threshold technique," *Journal of Medical Systems*, vol. 40, no. 4, p. 82, 2016.
- [53] B. Kurt, V. V. Nabiyev, and K. Turhan, "A novel automatic suspicious mass regions identification using Havrda & Charvat entropy and Otsu's N thresholding," *Comput Methods Programs Biomed*, vol. 114, no. 3, pp. 349–360, 2014.
- [54] R. Achanta and S. Süsstrunk, "Saliency detection using maximum symmetric surround," in *17th IEEE International Conference on Image Processing (ICIP)*, pp. 2653–2656, Hong Kong, September 2010.
- [55] N. Murray, M. Vanrell, X. Otazu, and C. A. Parraga, "Saliency estimation using a non-parametric low-level vision model," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 433–440, Colorado Springs, CO, USA, June 2011.
- [56] A. Oliver, J. Freixenet, J. Marti et al., "A review of automatic mass detection and segmentation in mammographic images," *Medical Image Analysis*, vol. 14, no. 2, pp. 87–110, 2010.
- [57] H. Cheng, X. Shi, R. Min, L. Hu, X. Cai, and H. Du, "Approaches for automated detection and classification of masses in mammograms," *Pattern Recognition*, vol. 39, no. 4, pp. 646–668, 2006.
- [58] N. M. Basheer and M. M. H. Mohammed, "Segmentation of breast masses in digital mammograms using adaptive median

filtering and texture analysis,” *International Journal of Recent Technology and Engineering*, vol. 2, no. 1, pp. 39–43, 2013.

- [59] L. Liu, J. Li, and Y. Wang, “Breast mass detection with kernelized supervised hashing,” in *International Conference on Biomedical Engineering and Informatics*, pp. 79–84, Shenyang, China, 2016.

## Research Article

# Organic Boundary Location Based on Color-Texture of Visual Perception in Wireless Capsule Endoscopy Video

Chengliang Wang <sup>1</sup>, Zhuo Luo <sup>1</sup>, Xiaoqi Liu,<sup>1</sup> Jianying Bai,<sup>2</sup> and Guobin Liao<sup>2</sup>

<sup>1</sup>College of Computer Science, Chongqing University, Chongqing 400044, China

<sup>2</sup>Department of Gastroenterology, Second Affiliated Hospital, Third Military Medical University, Chongqing, China

Correspondence should be addressed to Chengliang Wang; wangcl55@gmail.com

Received 7 July 2017; Revised 8 September 2017; Accepted 23 October 2017; Published 10 January 2018

Academic Editor: Yong Xia

Copyright © 2018 Chengliang 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 addresses the problem of automatically locating the boundary between the stomach and the small intestine (the pylorus) in wireless capsule endoscopy (WCE) video. For efficient image segmentation, the color-saliency region detection (CSD) method is developed for obtaining the potentially valid region of the frame (VROF). To improve the accuracy of locating the pylorus, we design the *Monitor-Judge* model. On the one hand, the color-texture fusion feature of visual perception (CTVP) is constructed by grey level cooccurrence matrix (GLCM) feature from the maximum moments of the phase congruency covariance and hue-saturation histogram feature in HSI color space. On the other hand, support vector machine (SVM) classifier with the CTVP feature is utilized to locate the pylorus. The experimental results on 30 real WCE videos demonstrate that the proposed location method outperforms the related valuable techniques.

## 1. Introduction

Wireless capsule endoscopy (WCE) was invented by a group of researchers in Baltimore in 1989 and introduced by Given Imaging Inc. as a commercial tool [1]. And it has a good performance in screening bleeding, ulceration, submucosal swelling, worms, polyps, and cancer, which is a vital breakthrough in the comprehensive examination of gastrointestinal tract (GI) which is painless [2, 3]. So far, the WCE system of Chongqing Jinshan Science and Technology Group (Jinshan) mainly consists of the camera capsule, image recorder, capsule guiding device, and image workstation [4]. The capsule enters the digestive tract from the mouth and captures the images of the digestive tract by a digital camera in the capsule. However, a WCE video generally lasts over 8 hours and contains more than 40,000 frames which bring the clinician a heavy burden for the diagnosis and treatment of diseases. Even an experienced clinician will take over an hour at least to analyze the data of each patient on average. Therefore, it is necessary to detect lesion images automatically [2]. But different digestive tract organs have

different textures, which means that the same lesion appears differently in different organs as shown in Figure 1. For this reason, the computer researchers [5, 6] generally find the organic boundaries for segmenting the WCE video according to different organs before recognizing lesion images in a single organ. Yuan and Meng [5] and Karargyris and Bourbakis [6] just detect the lesion images in the small intestine, but they both manually select the part of WCE video about the small intestine for avoiding the disturbance of other organs. However, it is a time-consuming and laborious task to locate the organic boundaries artificially. So it is important to automatically locate organic boundaries, which not only can help the clinician to ensure relevant organ section and reduce the review time, but also is the vital preliminary work for automatic lesion recognition [7]. In this paper, we propose a novel method to locate the boundary between the stomach and the small intestine (the pylorus).

Many works are for WCE video segmentation and have been devoted to locate the pylorus because the key of WCE video segmentation is finding the organic boundaries. Although the researches on this work are not mature enough,

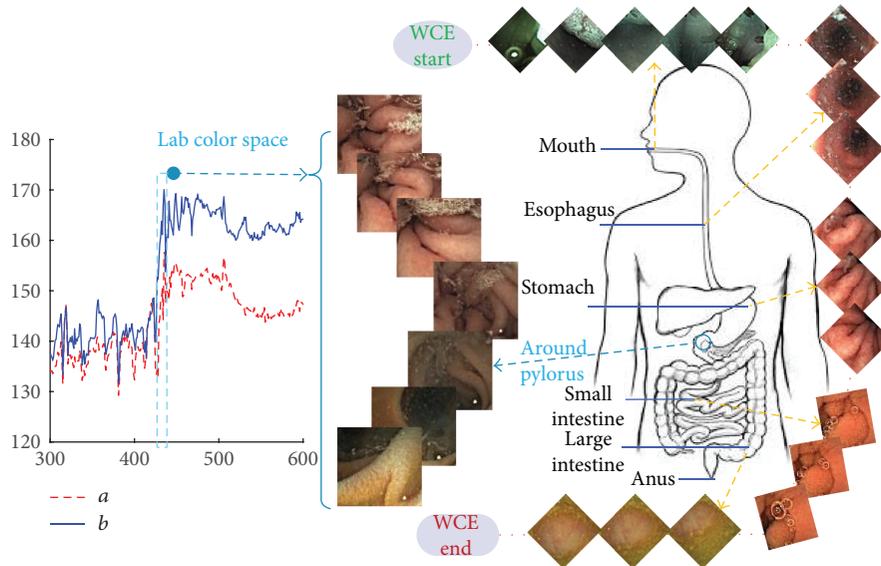


FIGURE 1: WCE videos about different organs of the digestive system and color change curve in Lab color space around the pylorus. WCE videos are provided by Jinshan. And some sample frames of a video show the results of screening digestive system by WCE. The pylorus is the end of the stomach and the boundary between the stomach and the small intestine, where the color of the images has changed markedly as shown in the  $a/b$  color curve. After images have been converted into Lab color space, the  $a/b$  color curve is generated by the average values in channel  $a$  and  $b$  of the continuous images around the pylorus.

they have yielded a great number of positive results. Cunha et al. [7] utilize MPEG-7 scalable color and SVM classifier to segment a WCE video into four parts including the entrance, stomach, small intestine, and large intestine based on Gaussian prior probabilities. In terms of their global model fitting step, this step is a time-consuming procedure for estimating and judging all frames in a WCE video. Some other classifiers based on color have been proposed to locate the pylorus. Berens et al. [8] report a stomach/intestine classifier with the hue-saturation histograms to predict the pylorus. However, this method just achieves an average performance of 86% and 85% for accuracy and recall, respectively. Li et al. [9] use color histogram in Lab color space and textural features in wavelet domain to represent the visual content. Furthermore, they apply motion analysis approaches to segment WCE video [10]. But the best average performance of CE video segmentation in terms of recall is 71.89%. Especially, the researchers all propose a two-level approach for WCE video segmentation [11, 12]. They all firstly find the approximate positions of organic boundary and then refine the boundary. However, these two approaches are both time-consuming tasks because they need to compute almost all the images about the stomach and the small intestine in WCE videos. Although Zhou et al. [11] considers the influence of the impurities, gastric juice, and illumination, they only use some fixed thresholds that weaken the robustness of denoting the valid regions of frames. And at the rough level, he computes a completed average dissimilarity curve to find the probable boundary, which adds some computing burden. Zhou et al. [12] applies the trained KNN classifier to the improved WLD features of the images around the candidates and selects three best candidates as the output in the end. However, this method needs

manual intervention to select the best one from the three candidate positions of the pylorus.

All of these above methods have ignored many interfering factors that lead to the dissatisfied accuracy and are time-consuming because they must deal with almost all the frames of the WCE video. The proposed method in this paper successfully shortens the time of locating the pylorus and improves the location accuracy by contrast with some of the above methods.

The contributions in this paper can be summarized into the following three points:

- (1) The CSD method is proposed to obtain the VROF region effectively and adaptively to remove the bad effects of the disturbances including food debris, strong shadows, overexposure, air bubbles, and gastric juice.
- (2) The *Monitor-Judge* model is designed to locate the organic boundary (the pylorus) for reducing time consumption.
- (3) The CTVP feature is constructed by grey level co-occurrence matrix (GLCM) feature from the maximum moments of the phase congruency covariance and hue-saturation histogram feature in HSI color space, which is better to express the difference between stomach images and small intestine images than other selected features [11, 12].

## 2. Materials and Methods

In this section, a novel method is proposed to locate the pylorus in this paper. Firstly, WCE video images are divided

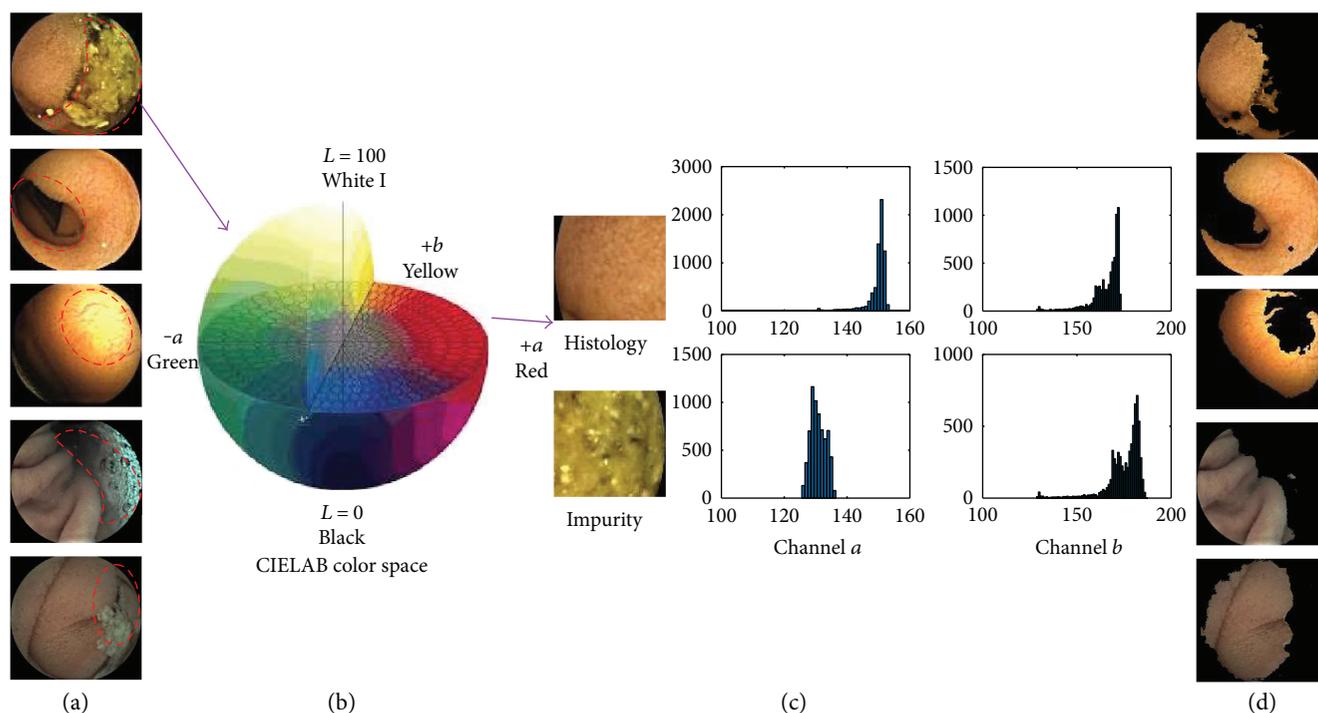


FIGURE 2: Process results of obtaining the VROF region and histogram comparison in Lab color space between digestive tract histology and the disturbance. The descriptions of the figure from left to right now follow. (a) Samples with five categories of the disturbances marked by red dashed line that are food debris, strong shadows, excessive bright region, air bubbles, and gastric juice consecutively. (b) Lab color model. (c) The comparison result of histograms between digestive tract histology and the disturbance in channels  $a$  and  $b$  in Lab domain. (d) Images processed by CSD method. The black regions are invalid and the colored parts are used for feature extraction.

into small windows that form a window pair sequence, and we obtain the VROF region of images by the CSD method. Secondly, we propose the *Monitor-Judge* model for locating the pylorus: *Monitor* constantly monitors the suspicious window pair with the pylorus according to the ratio of the color dissimilarity of current window pair and the average color dissimilarity of previous window pairs; *Judge* classifies images and estimates the pylorus position in the suspicious window pair by SVM classifier with the CTVF feature.

**2.1. VROF Region Extraction.** Actually, many possible disturbances such as gastric juice, shadows, excessive bright regions, and air bubbles show various appearances in different images and make it difficult to extract color or texture feature from the histology of digestive tract and therefore cause these extracted features unreliable.

The disturbances are as follows. (1) Food debris: food debris is one of the common impurities that obscure digestive tract tissue; (2) strong shadows: strong shadows are the lack of describing the real color and texture of digestive tract tissue; (3) overexposure: overexposure is one kind of image distortion caused by fierce reflection; (4) air bubbles: air bubbles are mainly caused by gastrointestinal peristalsis and pressure change; (5) and gastric juice: gastric juice is liquid commonly found in the stomach. Figure 2(a) gives some examples of WCE images with disturbances.

Valid region of the frame (VROF) is the region of a WCE image without any disturbance. VROF can show the color and texture of digestive tract tissue clearly.

CIE Lab color is designed to approximate human vision, which aspires to perceptual uniformity [13]. The three coordinates of CIELAB represent the lightness of the color ( $L^* = 0$  yields black and  $L^* = 100$  indicates diffuse white; specular white may be higher), its position between red/magenta and green ( $a^*$ , negative values indicate green while positive values indicate magenta), and its position between yellow and blue ( $b^*$ , negative values indicate blue and positive values indicate yellow). For removing these disturbances, we convert WCE images into Lab color space and find that tissue of digestive tract shows commonly stronger color reflection in channel  $a$  than the color reflection in any disturbance, and the tissue has obvious difference in channel  $b$  with any disturbance. Figure 2(c) shows the difference in histograms of channel  $a$  and channel  $b$  between digestive tract tissue and the disturbance. In terms of Lab color, digestive tract tissue in the image is salient region in contrast with disturbance region.

IG [14] as one of the saliency detection methods is simple to implement and computationally efficient. For obtaining the VROF region (salient region) in WCE images, the color-saliency detection (CSD) method is proposed in this paper based on IG. In the remainder of this subsection, details about the CSD method are described.

DoG filter, as a kind of band pass filters, is widely used for edge detection because it approximates the Laplacian of Gaussian (LoG) filter but it is much faster to compute than the LoG filter [14]. It satisfies detecting intensity changes when the standard deviations of the Gaussians are in the ratio 1:1.6 [14, 15]. DoG filter is chosen for WCE image

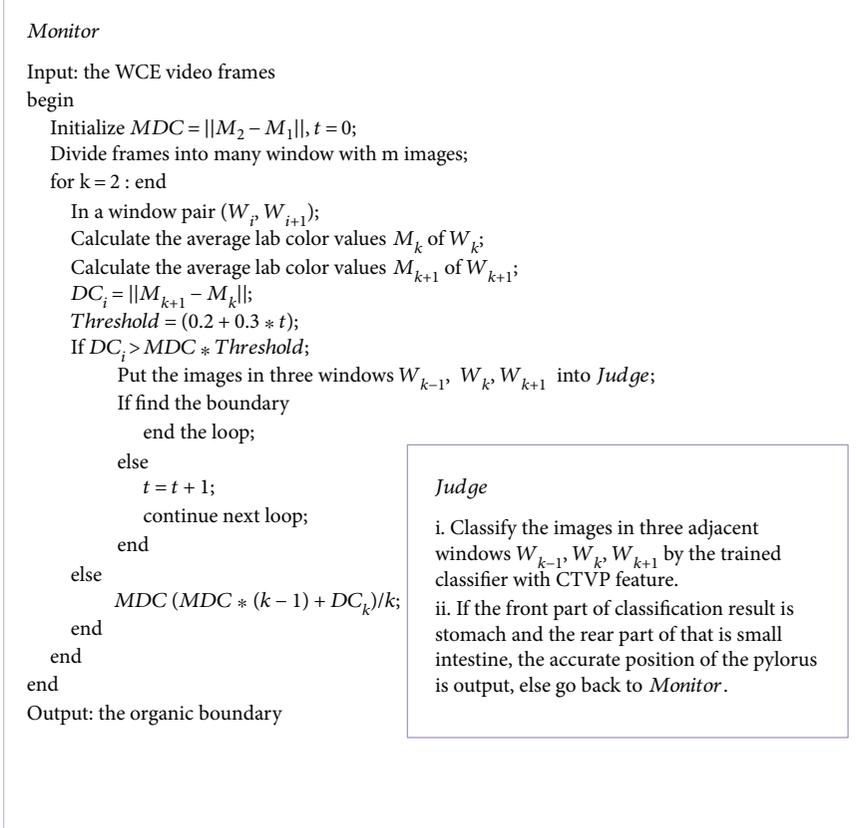
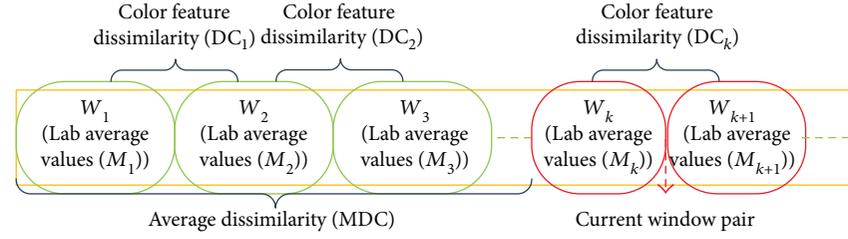


FIGURE 3: *Monitor-Judge* model. The procedures are described simply as follows. (1) Initialize  $t = 0$ ,  $MDC = DC_1$ ; (2) compute the average Lab color values  $M_k$  and  $M_{k+1}$  in the current window pair ( $W_k, W_{k+1}$ ); (3) compute the color dissimilarity  $DC_k$  in the window pair; (4) if  $DC_k / MDC > 0.2 + 0.3 * t$ , *Judge* checks the current window pair and  $t = t + 1$ , or go back to (2) to monitor the next window pair.

reprocessing in the CSD method. The DoG filter is given by the following:

$$DoG(x, y) = \frac{1}{2\pi} \left[ \frac{1}{\delta_1^2} e^{-(x^2+y^2)/2\delta_1^2} - \frac{1}{\delta_2^2} e^{-(x^2+y^2)/2\delta_2^2} \right]. \quad (1)$$

Our method of finding the saliency map  $S$  for an image of width  $W$  and height  $H$  pixels is formulated as follows:

$$S(x, y) = (a_u - a(x, y))^\alpha + (b_u - b(x, y))^\beta, \quad (2)$$

where  $a_u$  and  $b_u$  are the arithmetic mean pixel values of channels  $a$  and  $b$  in Lab color space.  $a(x, y)$  and  $b(x, y)$  are the corresponding image pixel values in the Gaussian blurred results of the original image in Lab color space and  $\alpha$  is greater than one and must be an odd number. When  $\alpha$  is greater, the saliency region represents the area with stronger value in channel  $a$ .  $\beta$  should be greater than one and smaller

than  $\alpha$ . The experiments show that  $\alpha = 3$  and  $\beta = 1.5$  contribute to good results presented in Figure 2(d).

VROF is the region of a frame where  $S(x, y) > 0$  and the results are presented in Figure 2(d).

## 2.2. Monitor-Judge Model for Pyloric Position

**2.2.1. Monitor-Judge Model.** In WCE videos with good quality, the sequence frames have an obvious color change when a capsule enters the next digestive organ. These characteristics can be found from the  $a/b$  color curve in Figure 1. Therefore, we design *Monitor-Judge* model for monitoring and judging the suspicious window pair with the pylorus as shown in Figure 3. For efficiency, we can divide the WCE video into many small windows with  $m$  images, which form window pair sequences  $(W_1, W_2), (W_2, W_3), \dots, (W_{n-1}, W_n)$ . In this method, images are converted into Lab color space. In order to reduce the influence of luminance, only the average values

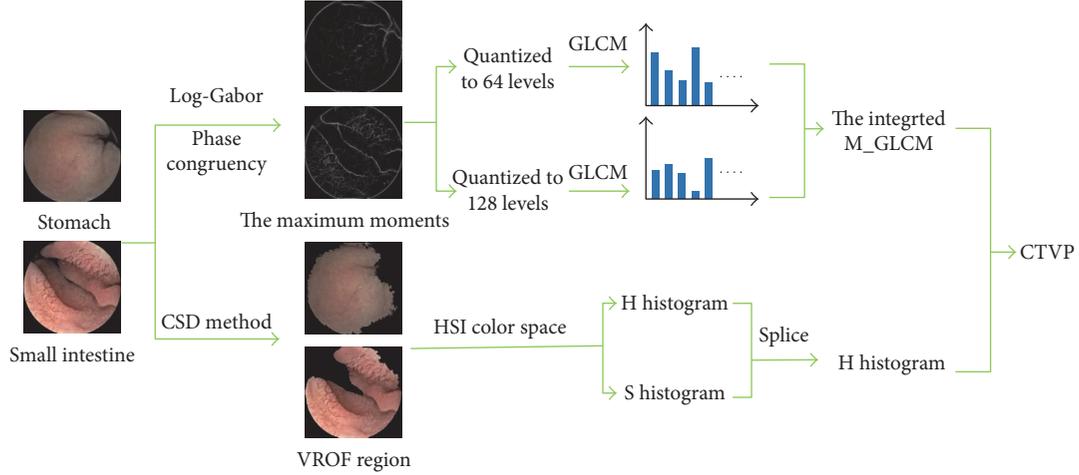


FIGURE 4: Construction of the fusion feature (CTVP) for describing the WCE image effectively.

of data in channels  $a$  and  $b$  of the VROF region are considered as colored features in this procedure.

$$\text{avg}_c = \frac{\sum_{i,j} I_c(i,j)}{N_{\text{valid}}}, \quad (3)$$

$$M_{c,k} = \frac{\sum_1^m \text{avg}_c}{m},$$

where  $\text{avg}_c$  and  $M_{c,k}$  represent the average color values of a frame and the image window  $k$  respectively;  $c$  presents any channel of Lab color space, that is,  $L$ ,  $a$ , or  $b$ .  $I_c(i,j)$  is the value of the pixel in the valid regions, and  $N_{\text{valid}}$  represents the total number of pixels in the valid region of a frame. Then, Euclidean distance is utilized to demonstrate the color dissimilarity of a window pair.

$$\text{DC}_k = \sqrt{(M_{a,k} - M_{a,k+1})^2 + (M_{b,k} - M_{b,k+1})^2}, \quad (4)$$

$$\text{MDC} = \frac{\sum_{t=1}^{k-2} \text{DC}_t}{k-2},$$

where  $\text{DC}_k$  stands for the color dissimilarity in  $(W_k, W_{k+1})$ . MDC stands for values in the average dissimilarity previous window pairs.

By analyzing many WCE videos, we find that the color of the WCE images about the same digestive organ generally have little change and the color of WCE images around the organic boundary has changed markedly. According to this, the key idea of the *Monitor-Judge* model is constantly comparing the color dissimilarity ( $\text{DC}_k$ ) of the current window pair  $k$  with the average color dissimilarity (MDC) of all previous window pairs and then considering detecting

organic boundary in the current window pair. Obviously, the proposed model does not need to deal with these images after finding organic boundary and it is more efficient than the two-level approaches in [11, 12]. The experiment shows that it just needs 2.55 times of *Monitor-Judge* operation on average to find the organic boundary.

In this model, the time complexity is  $O(m \cdot W \cdot H)$  which is better than the method in [11];  $m$  is the size of window and  $W$  and  $H$  are the width and height of an image, respectively. The pylorus appears in 400th~6000th images in our data. The intestinal peristalsis makes the capsule move forward slowly. The camera in the endoscopy takes three pictures per second, so that at least five continuous images are exactly similar. To improve efficiency, an interval of five frames in a window with 100 frames for extracting color feature will greatly reduce the cost of time through many experiments.

**2.2.2. The Color-Texture Fusion Feature of Visual Perception (CTVP).** Because there are some changes of color and texture between two adjacent organs in general, a classifier with the CTVP feature is applied to a target window pair shown in Figure 4.

The small intestine has a large number of small intestine villi as texture with contrast with the stomach, and the maximum moments can highlight the textures shown in Figure 5. To extract a useful texture feature, we firstly calculate the maximum moments of phase congruency [16] based on Fourier components of original image deal with 2D log-Gabor filter [17], and then extract the integrated GLCM features from the maximum moments.

The 2D phase congruency  $\text{PC}_{2D}(x)$  is defined as follows:

$$\text{PC}_{2D}(x) = \frac{\sum_S \sum_O W_o(x) \left[ A_{so}(x) \left[ \cos(\varphi_{so}(x) - \overline{\varphi_o(x)}) - \sin(\varphi_{so}(x) - \overline{\varphi_o(x)}) \right] - T_o \right]}{\sum_S \sum_O A_{so}(x) + \varepsilon}, \quad (5)$$

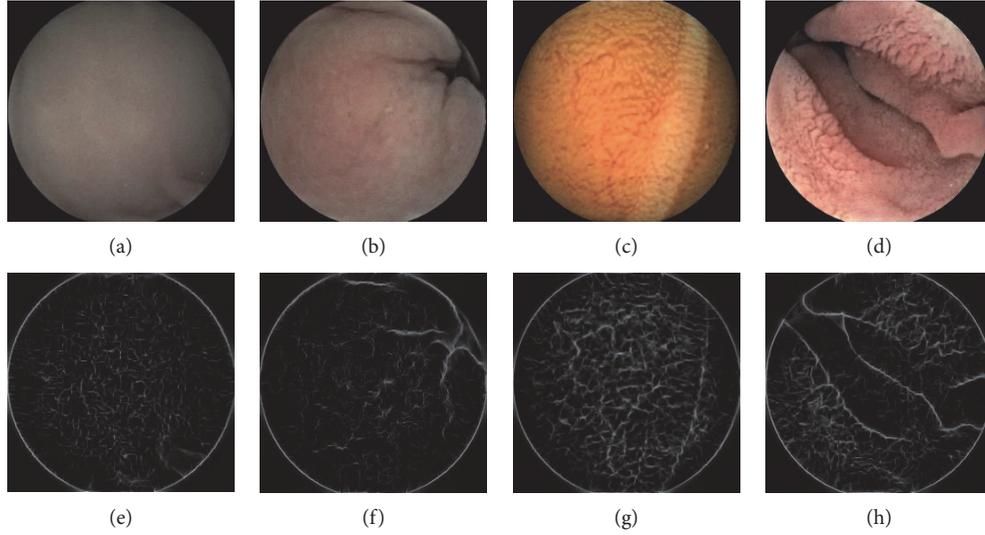


FIGURE 5: Comparison of the maximum moments of the phase congruency covariance between the stomach and the small intestine.

where the numerator is the weighted and noise compensated local energy summed over all orientations, and the denominator is the total sum of filter response amplitudes over all orientations and scales [16].  $x$  is the pixel location in the spatial domain.  $W_o(x)$  is weighing function of phase congruency by frequency spread at orientation  $o$ .  $A_{so}(x)$  denotes the amplitude of the grey scale WCE image.  $\varphi_{so}(x)$  denotes the phase response of original image at scale  $s$  and orientation  $o$  of log-Gabor filter.  $\overline{\varphi_o(x)}$  represents the mean phase angle at orientation  $o$ .  $T_o$  is the estimated noise energy at orientation  $o$ .  $\varepsilon$  is small constant which prevents division by zero.  $[ ]$  symbol denote that the enclosed quantity is equal to itself when its value is positive and zero otherwise. An overall measure of phase congruency in the two-dimensional (2D) local energy is firstly calculated in several orientations (typically six) by using data from oriented 2D Log-Gabor wavelets [17].

The maximum moments of the phase congruency covariance is given by

$$M = \frac{\left( a + c + \sqrt{b^2 + (a - c)^2} \right)}{2}, \quad (6)$$

where  $a = \sum (\text{PC}_{2D}(o) \cos(o))^2$ ,  $b = \sum \text{PC}_{2D}(o)^2 \sin(o) \cos(o)$ , and  $c = \sum (\text{PC}_{2D}(o) \sin(o))^2$ .  $\text{PC}_{2D}(o)$  refers to the phase congruency value determined at orientation  $o$  and the sum is performed over the discrete set of orientations.

After getting the map of maximum moment, the cooccurrence probabilities are calculated from this map as texture features. These probabilities represent the conditional joint probabilities of all pairwise combinations of grey levels in the spatial window of interest given by two parameters: inter pixel distance ( $\delta$ ) and orientation ( $\theta$ ) [18]. The probability measure can be defined as follows:

$$C_{i,j} = \frac{P_{ij}(\delta, \theta)}{\sum_{i,j} P_{ij}(\delta, \theta)}, \quad (7)$$

where  $P_{ij}(\delta, \theta)$  represents the number of occurrences of grey level pair  $(i, j)$  within the given window, given a certain  $(\delta, \theta)$  pair, and  $G_{i,j}$  is the quantized number of grey levels. In this paper, we consider four properties as features concluding *contrast*, *correlation*, *energy*, and *homogeneity* that are formulated as follows:

$$\begin{aligned} \text{Contrast} &= \sum C_{i,j} (i - j)^2, \\ \text{Correlation} &= \sum \frac{(i - u_x)(j - u_y) C_{i,j}}{\sigma_x \sigma_y}, \\ \text{Energy} &= \sum C_{i,j}^2, \\ \text{Homogeneity} &= \sum \frac{(i - u_x)(j - u_y) C_{i,j}}{1 + |i - j|}. \end{aligned} \quad (8)$$

By analyzing many WCE images in the stomach and the small intestine, it is found that color of images between the stomach and the intestine has obvious discrimination in general. It is reported that color histogram contributes to good performance of described images among different digestive organs [11]. HSI color space decomposes an image into components of hue (H), saturation (S), and intensity (I) [19]. The intensity of image is instable because of constant movement of capsule endoscopy. So, we also choose HS histogram to represent the color features. Importantly, after RGB images are converted into HSI space, we calculate the HS histogram from the VROF region.

$$\begin{aligned} \text{Hist}_H(i) &= \frac{n_H(i)}{N_{\text{valid}}}, \\ \text{Hist}_S(i) &= \frac{n_S(i)}{N_{\text{valid}}}, \end{aligned} \quad (9)$$

where  $N_{\text{valid}}$  is the total number of pixels in the VROF region,

$n_H(i)$  and  $n_S(i)$  are the frequency of the  $i$ th bin in H and I channels of the VROF region, and  $i = 1, 2, \dots, 16$ .

SVM classifier [20] with RBF kernel is utilized for classifying images in three target windows. Then we judge the position of the pylorus from the classification results. In this step, the time complexity from GLCM is  $O(\eta N)$  where  $\eta$  is the range of the intensity level (e.g., 256).

### 3. Results and Discussion

**3.1. Dataset and Experiment Design.** The WCE video data used in these experiments is acquired from different patients with different ages and provided by Jinshan in Chongqing. There are 30 videos being used for the experiments in this paper, and each video contains more than 42000 frames with  $256 \times 240$  pixels. It is worth mentioning that our samples are very diverse and come from all over the world like China, Middle East, and Europe. For privacy reasons, the names of the cases are not real and just indicate where they are from. We randomly select 3801 images around the pylorus in eight videos, and 1822 are before the pylorus, 1979 are after the pylorus. That is to say, there are 3801 frames included in the training and validation set. The label of the images that before and after the pylorus are 0 and 1, respectively. In addition, another nine cases are used for testing the accuracy of the estimate position of the pylorus based on the proposed method in this paper. The rest of the videos are for demonstrating the importance of the VROF region.

The accuracy of locating the pylorus is assessed by the error frames between the boundary obtained from the experiments and the one manually labelled by three clinicians. The mean and the median errors are considered in experiment results. The mean error is the average error of all test videos and the median error is the middle error value in all test videos. To verify the effectiveness of our proposed algorithms, three traditional performance metrics such as accuracy, sensitivity (recall), and specificity are measured in our experiments. Those three performance metrics are described as follows:

$$\begin{aligned} \text{Accuracy} &= \frac{P_T + N_T}{P_T + P_F + N_T + N_F}, \\ \text{Sensitivity} &= \frac{N_T}{P_F + N_T}, \\ \text{Specificity} &= \frac{P_T}{N_F + P_T}, \end{aligned} \quad (10)$$

where  $P_T$  is the number of actually positive frames predicted as positive frames,  $N_T$  is the number of actually negative frames predicted as negative frames,  $P_F$  is the number of true positive frames predicted as negative frames, and  $N_F$  is the number of true negative frames predicted as positive frames.

To demonstrate the operation of the proposed method, we perform six sets of experiments.

The first experiment confirms the importance of segmenting the VROF region. The second one compares the performance of SVM classifier, K-nearest neighbour (KNN) classifier, and Naive Bayes classifier. SVM classifier with

TABLE 1: Performance comparison of using entire frame and valid region.

	Frequency				Time consumption (s)
	Mean	Median	Max	Min	Average
Entire region	3.59	1.5	19	1	161
VROF region	2.55	1.5	7	1	16

Frequency: the times of monitoring needed to find the right window with the pylorus. Time consumption: time consumption for finding the right window with the pylorus.

RBF kernel is selected in this experiment. The parameter  $K$  used in KNN classifier is 5. The third one compares the CTVF feature with HS histogram, the ULBP feature, and the improved WLD. The fourth one evaluates the influence of different window sizes on locating the pylorus. The fifth one compares and analyzes the performance of different methods. The last one gives analysis of the location error of the proposed method.

In order to obtain convincing results, all the systems run on our own data. On the one hand, our results of the proposed method are compared with the existing methods which we try our best to reimplement based on the literature. On the other hand, these results are also compared with the ground truth (GT); GT has been generated artificially by three clinicians who are from Third Military Medical University, China. For convincing research, the error standard is set up by those three clinicians.

**3.2. Experiment Results.** The first experiment is for revealing the importance of segmenting the VROF region. Table 1 shows the efficiency of finding right window with the pylorus based on entire frames and VROF regions by frequency and time consumption. Frequency is smaller, the efficiency of finding the right window with the pylorus is better. Based on the VROF region, it just needs 2.55 times of the *Monitor-Judge* operation and about 16 seconds to find the right window with the pylorus. The results clearly confirm the significance of valid region denotation. And the better performance of using VROF regions can be explained by the fact that it is necessary to reduce the negative influence brought by gastric juice, shadows, excessive bright regions, and bubbles.

The second experiment is for evaluating the classification performance of different classifier and selecting the appropriate classifier in the proposed method. As shown in Figure 6, we apply 10-fold cross validation operations 10 times to evaluate the classification performance of KNN classifier, Naive Bayes classifier, and SVM classifier based on the different values of  $\alpha$  and  $\beta$  in the CSD method. It is found that the CTVF feature is reliable and effective to describe the WCE images of different digestive organ, because SVM classifier and KNN classifier both obtains the acceptable classification results. From Figure 6(a),  $\alpha = 3$  and  $\beta = 1.5$  contribute to the best performance of KNN classifier and SVM classifier; and the average accuracy of SVM classifier are 98.9%, which has an increase of 1% than that of KNN classifier. Because a peculiarity of the KNN classifier is that it is sensitive to the local structure of the training data, the SVM classifier with RBF kernel is a more appropriate choice for the proposed method.

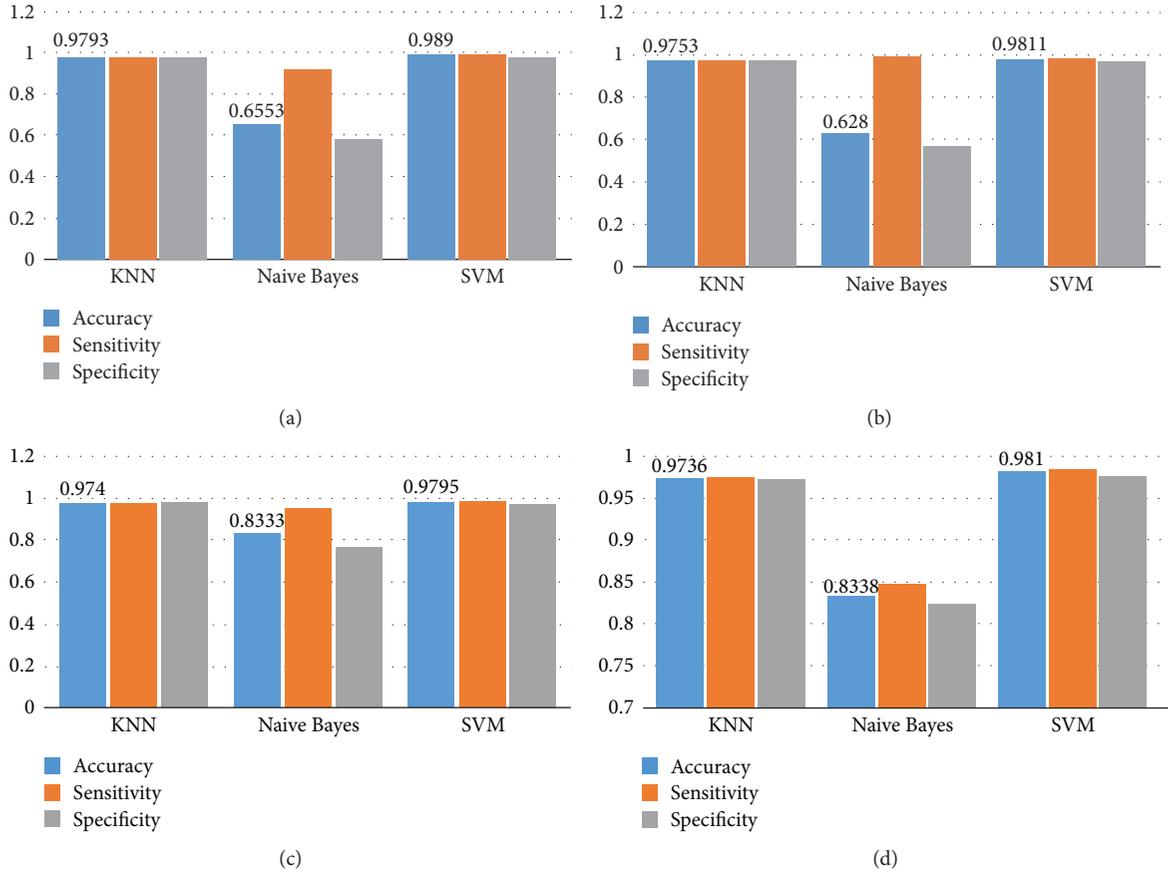


FIGURE 6: Classification performance of different classifiers and the influence of different parameters of the CSD method on the CTVP feature. (a) The CSD method with  $\alpha = 3$ ,  $\beta = 1.5$ . (b) The CSD method with  $\alpha = 5$ ,  $\beta = 3$ . (c) The CSD method with  $\alpha = 7$ ,  $\beta = 5$ . (d) CSD method with  $\alpha = 9$ ,  $\beta = 7$ .

TABLE 2: Classification rates of different features.

Features	Accuracy	Sensitivity	Specificity
HS histogram and ULBP [11]	97.1%	98.7%	95.4%
Improved WLD [12]	86.0%	90.2%	82.5%
The integrated M_GLCM	94.09%	94.68%	93.51%
HS histogram (VROF region)	93.44%	96.13%	90.86%
CTVP	98.9%	99.7%	98.1%

The third experiment is for presenting a comparison between different features. The results in Table 2 are acquired by 10-fold cross validation method to display the performance of different features including HS histogram, the ULBP feature, the improved WLD, the integrated M\_GLCM, and the CTVP feature. It is found that the proposed the CTVP feature is more reliable and effective for classifying the stomach and the small intestine than these comparison features. The improved WLD is not good choice to describe the WCE images in our training data because the average accuracy of SVM classifier with the improved WLD is just 86.0%. The average accuracy of the CTVP feature are 98.9%, which is more suitable to describe an image than

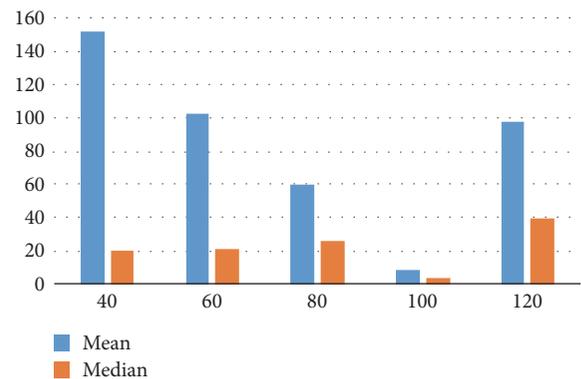


FIGURE 7: Mean error and median error of locating the pylorus in the proposed method with different sizes of sliding window. The  $x$ -axis represents the different sizes ( $m$  in Section 2.2.1) of sliding window and the  $y$ -axis shows the location error.

independent color or texture feature. HS histogram and ULBP [11] achieves an average performance of 97.1% and 95.4% for accuracy and specificity, respectively. Nevertheless, it is not more excellent than the CTVP feature in this paper.

TABLE 3: The results of the different location methods.

Cases	Clinicians' annotation	Zhou R.'s method [11]		Zhou S.'s method [12]		The proposed method	
	Position	Position	Time (min)	Position	Time (min)	Position	Time (min)
Duan	4849	5061		4861	3.89	4852	2.43
Liu	5440	4082		3921	2.88	5420	2.81
Feng	1634	4790		1271	3.41	1660	1.31
Wang	2042	1202	$\approx 141$	1991	1.77	2044	1.6
AMTA	710	578		501	2.86	708	0.56
AMM	1380	1192		1331	2.87	1380	0.37
RG	4019	1355		3531	2.33	4032	0.58
NNY	1704	4239		1511	3.23	1708	1.44
<i>Mean</i>	—	1385	141	360	2.91	9	1.39
<i>Median</i>	—	1099	141	201	2.78	4	1.26

*Mean* is the average error of locating the pylorus and *median* is the median of errors in the testing cases. *Time* represents time needed for locating the pylorus.

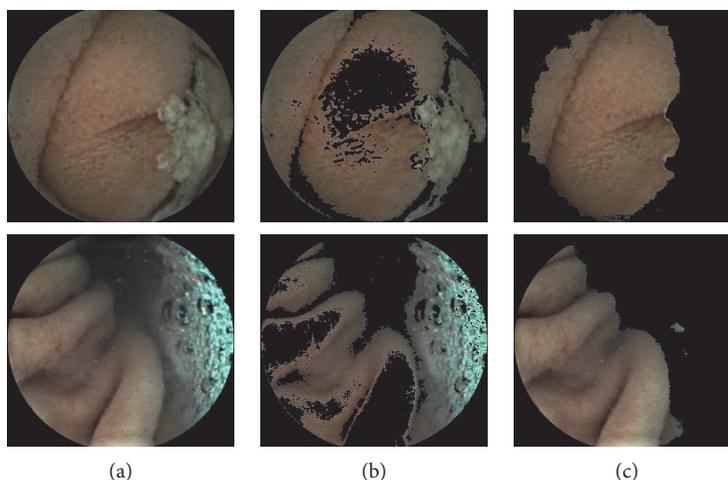


FIGURE 8: Different method for extracting the VROF region. (a) Original images with gastric juice and bubbles. (b) The process results of (a) based on Zhou R.'s operation [11]. (c) The process results of (a) based on CSD method.

The fourth experiment is for evaluating the influence of different window sizes on locating the pylorus. In Figure 7, we evaluate the influence of different values of  $m$  on locating the pylorus by mean error and median error. The results show that the proposed method with  $m = 100$  has the best performance of locating the pylorus. This is explained by the fact that  $m = 100$  can weaken the influence of classification error in our data set.

The fifth experiment is for comparing the performance of the proposed method with two existing valuable methods. In Table 3, we compare and analyze the performance of different location methods by the location accuracy and the time consumption. Zhou et al. [12] get three candidate positions of the pylorus as the output. Table 3 just shows the best one of three candidates of the pylorus based on Zhou S.'s approach. The error of Zhou S.'s approach is 360 and 201 frames of mean and median, respectively. It is because Zhou et al. [12] do not think about the disturbances in images, which makes a bad influence on the selection of the candidates of the pylorus and leads to poor performance of

classification. Although Zhou et al. [11] has considered the negative effects of the disturbances that cause the result of extracting features unreliable, it is lack of robustness to denote valid regions with some fixed condition in HSI color space. The examples in Figure 8 show that the proposed CSD method is more adaptive to extract the VROF region than Zhou R.'s method. Zhou R. uses the peak of the completed dissimilarity curve in Lab color space to locating the pylorus approximately. However, not only computing this completed dissimilarity curve is time-consuming, but also the position of the pylorus may not appear near the peak in our data because of the shortage of his mechanism for denoting the valid regions in WCE images. If they fail in the rough level, it is impossible to find the correct position of the pylorus in the fine level. These defects of Zhou R.'s method lead to the low accuracy (mean error is 1385 frames) and long execution time (141 minutes). The *Monitor-Judge* model shows excellent performance on locating the pylorus in WCE video: On the one hand, the error of locating the pylorus is just 9 and 4 frames of mean and median, respectively; on the

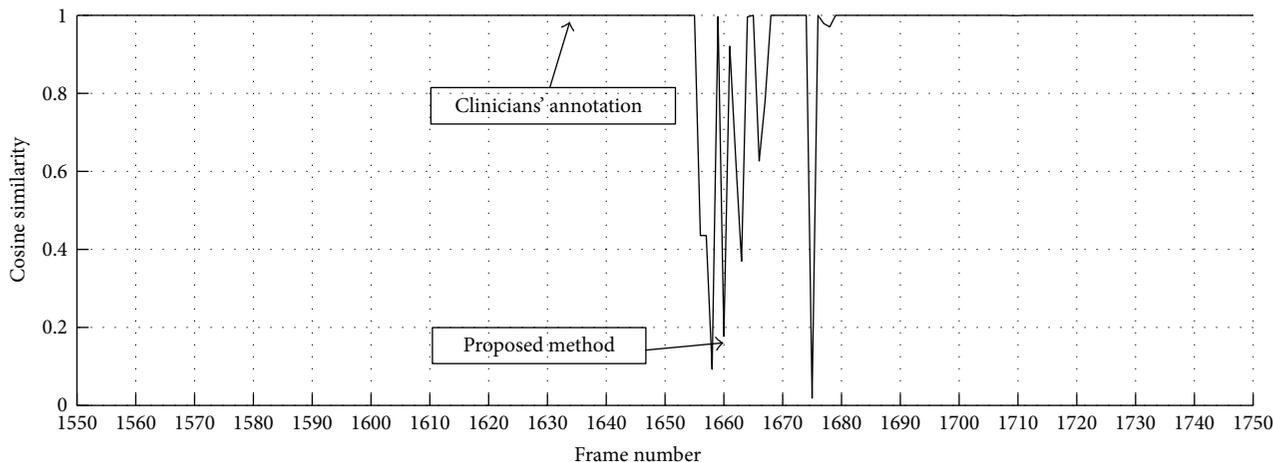


FIGURE 9: The curve of cosine similarity based on H histogram in HSI color space. The color features are extracted from 1551~1750 frames in the case—Feng. The 1634th frame is annotated as the pylorus by three clinicians and 1660th frame is the result of the proposed method.

other hand, the proposed method takes less time (1.26 minutes on average) to locating the pylorus than these two existing methods.

The sixth experiment is for analyzing the reason of the location error. As shown in Table 3, there is an obvious distinction of the position of the pylorus between clinician and our method in Feng. In Figure 9, we analyze the location error in the video (Feng) by the cosine similarity of color feature. The result shows that the color features of the frames around the position of the pylorus that annotated by clinician do not have a distinct change in general. However, the position that annotated by the proposed method is clearly marked boundary of the color feature. Due to some hardware problems or other unknown factors, this error is caused by the color distortion of the image that also makes trouble for the clinicians. In consideration of the importance of color feature for describing the stomach and the small intestine, the error is generally acceptable by the clinicians.

#### 4. Conclusions

In this paper, we have introduced an effective method to locate the pylorus in WCE videos. Firstly, the CSD method is designed to obtain the VROF region for the color feature extraction, which can remove the bad effects of the disturbances including food debris, strong shadows, overexposure, air bubbles, and gastric juice. Secondly, the *Monitor-Judge* model and the CTVF feature are proposed to promote the efficiency and accuracy of locating the pylorus. Based on color variation rule of sequential images, the proposed *Monitor-Judge* model reduces the time consumption of locating the pylorus. And the CTVF feature is very suitable to describe the images about the stomach and the small intestine. Thirdly, the SVM classifier is applied to predict frames in the target windows. And the position of the pylorus in WCE video is determined by analyzing the classification errors of the SVM classifier. Experimental results show that the proposed approach outperforms the techniques proposed in [11, 12] in terms of the location efficiency and accuracy on

a database of 30 WCE videos. In the future, we will collect more WCE videos to verify the validity of the method for locating the boundary between the small intestine and the large intestine and investigate new methods for abnormality detection in different digestive organs.

#### Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

#### Acknowledgments

The authors are grateful to the clinicians from Xinqiao Hospital, China, for their helpful discussion and suggestions. This work is supported by the National Natural Science Foundation of China under Grant no. 61672115 and Chongqing Social Undertakings and Livelihood Security Science and Technology Innovation Project Special Program (no. cstc2015jcyjBX0124 and no. cstc2017shmsA30003).

#### References

- [1] O. Lin, R. A. Kozarek, D. Schembre, and W. Spaulding, "Blinded comparison of esophageal capsule endoscopy (ECE) versus conventional esophagogastroduodenoscopy (EGD) for identification of esophagitis and Barrett's esophagus in patients with chronic gastroesophageal reflux disease (GERD)," in *Proceedings of the 4th International Conference on Capsule Endoscopy*, p. AB87, Maimi, FL. Yoqneam, Israel: Given Imaging, 2005.
- [2] S. Iogagiu, L. Ciobanu, and O. Pascu, "Colon capsule endoscopy: a new method of investigating the large bowel," *Journal of Gastrointestinal and Liver Diseases*, vol. 17, pp. 347–352, 2008.
- [3] M. Appleyard, Z. Fireman, A. Glukhovskiy et al., "A randomized trial comparing wireless capsule endoscopy with push enteroscopy for the detection of small-bowel lesions," *Gastroenterology*, vol. 119, no. 6, pp. 1431–1438, 2000.

- [4] “Capsule endoscopy - OMOM, Jinshan Science & Technology Inc,” <http://english.jinshangroup.com/>.
- [5] Y. Yuan and Q. H. Meng, “Polyp classification based on Bag of Features and saliency in wireless capsule endoscopy,” in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3930–3935, Hong Kong, China, 2014.
- [6] A. Karargyris and N. Bourbakis, “Detection of small bowel polyps and ulcers in wireless capsule endoscopy videos,” *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 10, pp. 2777–2786, 2011.
- [7] J. P. Silva Cunha, M. Coimbra, P. Campos, and J. M. Soares, “Automated topographic segmentation and transit time estimation in endoscopic capsule exams,” *IEEE Transactions Medical Imaging*, vol. 27, no. 1, pp. 19–27, 2008.
- [8] J. Berens, M. Mackiewicz, and G. D. Bell, “Stomach, intestine and colon tissue discriminators for wireless capsule endoscopy (WCE) images,” in *Proceedings of SPIE, 5747, Medical Imaging 2005, Image Processing*, no. 1–3, pp. 283–290, San Diego, CA, USA, 2005.
- [9] B. Li and M. Q.-H. Meng, “Capsule endoscopy video boundary detection,” in *2011 IEEE International Conference on Information and Automation*, Shenzhen, China, 2011.
- [10] B. Li, M. Q.-H. Meng, and C. Hu, “Motion analysis for capsule endoscopy video segmentation,” in *2011 IEEE International Conference on Automation and Logistics (ICAL)*, Chongqing, China, 2011.
- [11] R. Zhou, B. Li, Z. Sun, C. Hu, and M. Q.-H. Meng, “Wireless capsule endoscopy video automatic segmentation,” in *2012 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pp. 825–830, Guangzhou, China, 2012.
- [12] S. Zhou, H. Yang, M. A. Siddique, J. Xu, and P. Zhou, “A novel method for automatically locating the pylorus in the wireless capsule endoscopy,” *Biomedical Engineering Biomedizinische Technik*, vol. 62, no. 1, 2017.
- [13] G. Sharma and H. J. Trussel, “Digital color imaging,” *IEEE Transactions on Image Processing*, vol. 6, no. 7, pp. 901–932, 1997.
- [14] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk, “Frequency-tuned salient region detection,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1597–1604, Miami, FL, USA, 2009.
- [15] D. Marr, *Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information*, W. H. Freeman, San Francisco, CA, USA, 1982.
- [16] P. Kovese, *Phase Congruency Detects Corners and Edges*, DICTA, 2003.
- [17] J. Li, N. Sang, and C. Gao, *Log-Gabor Weber Descriptor for Face Recognition. Computer Vision - ACCV 2014 Workshops*, vol. 2014, Springer International Publishing.
- [18] R. M. Haralick, K. Shanmugam, and I. h. Dinstein, “Textural features for image classification,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-3, no. 6, pp. 610–621, 1973.
- [19] E. Welch, R. Moorhead, and J. K. Owens, “Image processing using the HSI color space,” in *Southeastcon '91. IEEE Proceedings of*, vol. 2, pp. 722–725, Williamsburg, VA, USA, 1991.
- [20] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.