Hindawi
Security and Communication Networks
Volume 2021, Article ID 7812223, 11 pages
https://doi.org/10.1155/2021/7812223

Research Article

A Novel Prediction Method Based on Artificial Intelligence and Internet of Things for Detecting Coronavirus Disease (COVID-19)

Shenqi Jing,[1,2,3,4] Qijie Qian,[5] Hao She,[5] Tao Shan,[1] Shan Lu,[6] Yongan Guo,[5] and Yun Liu[2]

[1] *School of Information Management, Nanjing University, Nanjing 210023, China*
[2] *Institute of Medical Informatics and Management, Nanjing Medical University, Nanjing 210023, China*
[3] *School of Biomedical Engineering and Informatics, Nanjing Medical University, Nanjing 210023, China*
[4] *Center for Data Management, The First Affiliated Hospital of Nanjing Medical University (Jiangsu Province Hospital), Nanjing 210023, China*
[5] *Engineering Research Center of Health Service System Based on Ubiquitous Wireless Networks, Ministry of Education, Nanjing University of Posts and Telecommunications, Nanjing 210023, China*
[6] *Department of Geriatrics, The First Affiliated Hospital of Nanjing Medical University (Jiangsu Province Hospital), Nanjing 210023, China*

Correspondence should be addressed to Yongan Guo; guo@njupt.edu.cn

Received 15 June 2021; Revised 14 August 2021; Accepted 2 November 2021; Published 27 November 2021

Academic Editor: Qingqi Pei

Novel coronavirus spreads fast and has a huge impact on the whole world. In light of the spread of novel coronaviruses, we develop one big data prediction model of novel coronavirus epidemic in the context of intelligent medical treatment, taking into account all factors of infection and death and implementing emerging technologies, such as the Internet of Things (IoT) and machine learning. Based on the different application characteristics of various machine learning algorithms in the medical field, we propose one artificial intelligence prediction model based on random forest. Considering the loose coupling between the data preparation stage and the model training stage, such as data collection and data cleaning in the early stage, we adopt the IoT platform technology to integrate the data collection, data cleaning, machine learning training model, and front- and back-end frameworks to ensure the tight coupling of each module. To validate the proposed prediction model, we perform the evaluation work. In addition, the performance of the prediction model is analyzed to ensure the information accuracy of the prediction platform.

1. Introduction

According to the statistics of the coronavirus disease 2019 (COVID-19) pandemic reported by the World Health Organization (WHO), there are already more than 56 million confirmed cases and 1.35 million deaths as of 15 : 59 Central European Time (CET),November 20, 2020, which indicate a very serious global epidemic situation. The number of COVID-19-infected patients has exceeded 1 million in many countries, including the United States, India, Brazil, and France. The United States, in particular, has over 10 million confirmed cases of COVID-19. Thus, it is critical to conduct a status analysis and research on the effects of epidemic prevention and control measures based on different epidemic situations in countries worldwide.

Mathematical models are often used by researchers to derive the nonspreading conditions of infectious diseases and predict and analyze the trends of epidemics and infected populations. Correspondingly, relevant strategies are developed accordingly. One of the currently used epidemic prediction models is the Malthusian growth model[1]. However, this model still has a long way to go before being applied in the real world. A logistic regression model[2], also known as the SI model ($S =$ suspect, $I =$ infected), has been proposed to distinguish between infected and uninfected individuals. The SI model's predictions are also unrealistic when cure factors are not taken into account. Furthermore, the SIS model (S = suspect, I = infected, and R = recovered) was learned by comparing the behaviors in different regions. We propose the SEIR epidemic models (S = suspect, E = exposed,I = infected,and R = recovered) to predict when the cured population without immunity is more vulnerable to reinfections. On the contrary, the classic SIR model is commonly selected for a cured population with strong immunity. This classic SIR model is widely used to describe the overall trends of epidemics owing to its ease of operation and clear and concise structure. This classic SIR model was used, for example, to analyze the 2003 SARS epidemic. The epidemic's evolution and the overall spread patterns of the disease are described[3]. Based on this, the SEIR model introduces the exposed, that is, class E population, which considers that only part of the people who are easily infected and had contact with infected people are infectious, which makes the transmission cycle of the disease longer. However, more detailed factors are not considered in the process of epidemic prediction.

To complete the task of epidemic prediction, we developed a novel prediction method based on artificial intelligence (AI) and the Internet of Things. On account of the existing model, many Internet machine learning algorithms can be employed for the prediction method. As a result, we investigated a research-related work and analyzed the common algorithms used in medical prediction. We aimed to find an optimal algorithm with good convergence characteristics and efficiency to complete the prediction. A new IoT platform for epidemic prediction was built using existing models. Relevant experiments were conducted, and the algorithm's benefits were also validated.

The remainder of this paper is organized as follows. We discuss the related work inSection 2and the algorithm model inSection 3. InSection 4, we present the design of the prediction platform. We perform the simulation work inSection 5. Finally, we conclude the paper.

2. RelatedWork

COVID-19 now has no real control over the world. There are about 170 million newly diagnosed cases in the world. Previously, there were 445539 newly diagnosed cases in a single day. The total number of deaths in the world is about 3500000, with the number of new deaths in a single day being 10000. There are 28 countries or regions in the world with more than 1 million newly diagnosed cases.

The epidemic model is a result of research on vaccination against smallpox in a paper presented by Daniel Bernoulli in 1760. The mathematical model research began in the early twentieth century. When Kermack and McKendrick studied the Black Death epidemic in London in 1927, they proposed the SIR compartment model. In the analysis of infectious diseases, the SIR mathematical model has preconditions. First, it considers population birth and death, which may have an impact on population size, but the impact is minimal. Second, the SIR model assumes that the susceptible and the infected populations have certain mobility, and the susceptible population will migrate to the infected population by a certain factor. Finally, the SIR model assumes that the infected population will enter the emigrant population with a fixed proportion coefficient, and the state is irreversible. SIR is an effective simulation model of infectious diseases. By dividing the population structure into three groups, namely, susceptible, infected, and displaced, we can simplify the transmission law of infectious diseases and obtain a more accurate transmission law of infectious diseases.

The traditional SIR epidemic prediction model divides the population into three categories: those who are not sick but are likely to be infected (S), those who have been infected and can infect (I), and those who have been healed or died (R). However, the factors affecting population dynamics such as birth, death, and migration are not considered. The population is a constant.

Compared with the SIR model, the SEIR model introduces the incubation period and adds the exposed who are in this latent period of infection. The healthy person who has come into contact with the patient does not become ill right away, but as the pathogen's carrier, he or she becomes E. This mechanism is very consistent with the novel coronavirus prediction.

Various teams at home and abroad have researched on epidemic trend prediction using statistical model, SEIR model improved by SEIR model, and machine learning model; however, the prediction results have large fluctuations. Scholars used the SEIR model to predict the inflection point and peak value. The classic SEIR model can be applied to any type of epidemic situation, but the infection of personnel flow must be taken into account. The modified SEIR model was then used to fit and analyze the prediction of COVID-19. However, a significant difference was observed between the prediction result and the number of people reported by the National Health Commission because the impact of the prevention and control measures on the flow of people was not taken into account. Based on existing epidemic prevention and control measures, industry insiders incorporated the flow of people into the SEIR model to forecast the epidemic situation and also concluded the effectiveness of the tourism ban. Researchers from Southeast University published a paper on medRxiv to evaluate the epidemic trend and risk of the COVID-19 outbreak by using the modified SEIR model. While there are many uncertain factors in the personnel network, both the SEIR and modified SEIR models must analyze a large number of parameters, including R0 and removal rate. In this regard, many researchers use machine learning methods for data prediction and random forest methods to divide different cities into different prevention and control levels, which provides a good reference for epidemic prevention and control[4].

At the same time, foreign scholars have conducted much research on the epidemic situation. The main research aim and findings of this paper were as follows:

(1) This paper aimed to examine and compare existing machine learning algorithms in the medical field and find the best algorithm model based on AI for

creating a prediction platform to prevent and control epidemics.

(2) This paper aimed to combine machine learning with IoT, a platform of the IoT for the prevention and control of COVID-19, based on the traditional epidemic model.

(3) This paper aimed to introduce the designed COVID-19 prediction platform in the order of data collection, data cleaning, machine learning training model, and front- and back-end frameworks.

(4) The simulation results show that the predictions made by the AI designed in this paper's random forest model are more accurate than those made by the logistic regression and support vector machine (SVM) algorithms.

3. Algorithm Model Analysis

This section reviews and compares the existing machine learning algorithms used in the medical field, intending to identify the most suitable algorithm model for creating a prediction platform to prevent and control epidemics.

3.1. LogisticRegression. Logistic regression is a classification algorithm commonly used to solve binary classification problems. This algorithm has been extensively studied in the industrial and medical fields because of not only its simplicity but also its strong interpretability. The essence of logistic regression is to use maximum likelihood estimation to approximate the parameters of a given distribution[5]. To date, logistic regression has been applied in many fields, in which there are many medical scenarios[6]. Unfortunately, this algorithm cannot solve the nonlinear problems in the medical field, although it has broad prospects for health care and has been greatly studied in disease diagnosis. Typically, epidemic predictions are not linear, thus limiting the use of the logistic regression algorithm in epidemic prediction. Due to the simplicity of its form, the accuracy rate cannot be guaranteed, which makes fitting true data distribution using the logistic regression algorithm difficult. As a result, the logistic regression algorithm is not recommended for predicting epidemics[7].

3.1.1. Support Vector Machines. Before the widespread application of deep learning algorithms as machine learning algorithms, SVMs were considered the optimal method for small-sample classification problems. SVM is a non-clustering technique that can calculate distances outside of the plane. When the specific parameters of a given model are assigned by a training set, SVM classification tasks only respond to the support vectors and have no relation with the data dimensions. The computational time and storage requirements are reduced as a subset of the sample set[8]. SVM is a classifier, and the maximum intervals between different cases are its primary indicators. The hyperplane positions can be used to obtain constraints. In other words, SVM is mainly used to ensure the correct classification of two data types in the future. However, the constraint requires a maximum distance between the classification line and that point within the maximum acceptable error. Furthermore, SVM is a kernel-based technique that allows for higher-dimensional space conversion using kernel functions. The structure of SVM hyperplane solutions allows them to solve quadratic programming problems while meeting the requirements of duality and convexity. In fact, for solving the convex optimization problem, the optimal plane can be determined according to strong duality[9].

3.2. Random Forest. A decision tree using the object attributes and relationships to build tree-like diagrams that can be obtained using probability analysis. Each branch of the decision tree represents a prediction direction in those diagrams, and each leaf node denotes the final prediction result. The decision tree prediction model classifies data into different classes using different object categories, allowing decision-related information to be intuitively displayed in the decision-making process[10]. On the contrary, a single decision tree is prone to overfitting, resulting in poor generalization ability. When a decision tree performs feature selection, the classification is accomplished based on the most suitable feature[11]. As already known, the key to most classification decisions should be based on a set of features rather than a particular feature. As a result, when dealing with epidemic data with characteristics such as large-scale and multiple features, the classic decision tree algorithm is ineffective[12]. Random forest was proposed in 2001 and has since been widely employed in classification and regression. To improve the classification efficiency, random forest creates the model using the bagging method combined with decision trees[13]. Because each decision tree is trained using a set of independent random samples and employing random attribute selection, there are no correlations between them. When a new sample is needed to make a decision, every tree in the forest is voted on, and the most voted type is selected as the sample type[14].Figure 1presents the random forest classification method.

Assuming that the model output vector length is 3, the probability vectors can be first obtained by training multiple sets of decision trees of the model, and the final output can then be determined by averaging multiple sets of the probability values. In regression problems, the random forest algorithm can also be employed. However, the overfitting of the decision tree algorithm can be a barrier. By increasing the number of decision trees, random forest prevents model overfitting. Meanwhile, random forest classifiers can deal with missing data, making the method suitable for the analysis of big data in epidemic environments where data collection is difficult. Therefore, the epidemic prediction model designed in this paper is based on the random forest algorithm.

As presented inTable 1, SVM is limited to small cluster samples, and its efficiency is low when there are too many observation samples. The logistic regression algorithm is sensitive to the multicollinearity of the model's independent variables. To reduce the correlation between candidate

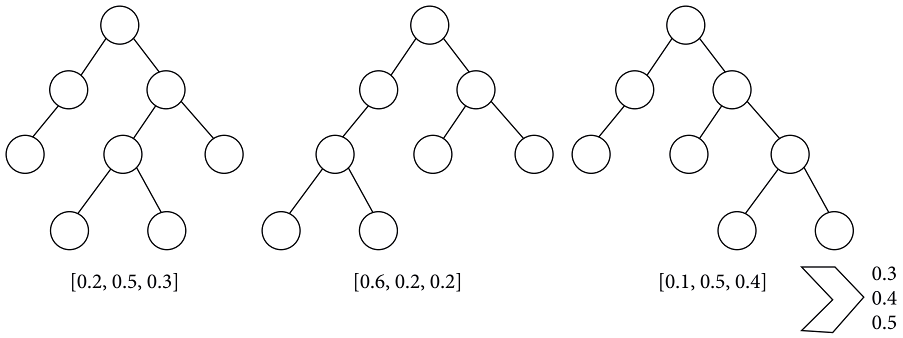

Figure 1: Illustration diagram of the random forest decision-making method.

Table 1: Algorithm analysis.

Algorithm name	Advantage	Shortcoming
Logistic regression	Simplicity, interpretability, widely used	Only deal with linear problems, unstable accuracy
Support vector machines	Good at small-sample processing, constraints can be obtained from the hyperplane positions	Large-scale samples are difficult to implement; multiclassification problem is difficult
Random forest	Samples with missing data can be processed, decision tree can be added to deal with overfitting	Overfitting problem

variables, it is important to select representative independent variables using factor analysis or variable cluster analysis. Random forest can process high-dimensional data (i.e., data with many characteristics) and does not require feature selection. It has excellent anti-interference and overfitting capabilities. In conclusion, for the input of epidemic data of a large order of magnitude, random forest algorithm is the most suitable.

Random forest algorithm is a supervised learning algorithm that uses an algorithm called bagging to combine many decision trees and classifies by voting mechanism. It has the advantages of fast training speed, strong generalization ability, and good classification performance. The following introduces the decision tree first and then the random forest algorithm next as the mathematical foundation of this article.

(1) Decision trees: Decision trees, also known as classification and regression trees (CART), can be used to describe different classes or values output after inputting a set of features. A decision tree is an example of a tree structure. Each internal node, branch, and leaf node denotes a different attribute test, test output, or final test result. Suppose that $\mathbf{X}$ is an input vector containing m features and is the output value, S is a training set containing n observations $(\mathbf{X}_i, \mathbf{Y}_i)$, where

$$S_n = \{(\mathbf{X}_1, \mathbf{Y}_1), \ldots, (\mathbf{X}_n, \mathbf{Y}_n)\}, \quad \mathbf{X} \in \mathbf{R}^m, \ \mathbf{Y} \in \mathbf{R}. \tag{1}$$

In the training process, the algorithm divides the input on each node. First, the CART algorithm recursively divides the input space $\mathbf{X}$ into two different branches:

$$\left\{\mathbf{X}^j < d\right\} \cup \left\{\mathbf{X}^j > d\right\}, \quad j \in \{1, \ldots, m\}, \ d \in \mathbf{R}. \tag{2}$$

For better division, (j, d)should minimize the cost function, usually the variance of the child nodes. The variance of the node$p$is defined as follows:

$$\mathrm{Var}(p) = \sum_{i:\ \mathbf{X}_i \in p} \left(\mathbf{Y}_i - \overline{\mathbf{Y}}_p\right)^2, \tag{3}$$

where $\overline{\mathbf{Y}}_p$ denotes the mean value of $\mathbf{Y}_i$ in node p and then divides the child nodes in the same way. The tree will stop when the maximum number of levels is reached or the number of observations contained in a node falls below a predetermined number. At the end of the training, a prediction function $\widehat{h}(\mathbf{X}, S_n)$ based on S_n will be established:

$$\widehat{\mathbf{Y}} = \widehat{h}(\mathbf{S}_n). \tag{4}$$

(2) Random forest: The random forest algorithm uses the Bootstrap sampling method to extract multiple samples from the original samples. It creates a decision tree model based on each Bootstrap sample. The predictions of multiple decision trees are then combined, and the final result is determined by voting. Random forest regression is a strong predictor that incorporates many weak predictors. Randomly select n replaced observation data from the original dataset S_n to obtain a Bootstrap sample. The random forest algorithm selects several Bootstrap subdatasets $(S_n^1, \ldots, S_n^q)$, then applies CART to these subdatasets, constructs some trees, and obtains a prediction function, such as (4).

Suppose that the training set is drawn from the independent and identically distributed random vectors $(\mathbf{X}, \mathbf{Y})$, $\mathbf{X}$ denotes the input vector, and $\mathbf{Y}$ represents the output vector; then, the mean square generalization error of the predicted output $h(\mathbf{X})$ is as follows:

$$E_{\mathbf{X},\mathbf{Y}}[\mathbf{Y}-h(\mathbf{X})]^2. \tag{5}$$

The prediction output of random forest regression is obtained by averaging h decision trees $\{h(\theta,\mathbf{X}_k)\}$, which has the following theorem.

Theorem 1. *Whenk* $\longrightarrow\infty$,

$$E_{\mathbf{X},\mathbf{Y}}\left[\mathbf{Y}-\overline{h}_k(\mathbf{X},\boldsymbol{\theta}_k)\right]^2 \longrightarrow E_{\mathbf{X},\mathbf{Y}}\left[\mathbf{Y}-E_{\boldsymbol{\theta}}(\mathbf{X},\boldsymbol{\theta}_k)\right]^2. \tag{6}$$

Mark the right part of (6) as PE^*, which is the generalization error of the random forest. The average generalization error PE of each decision tree can be defined as follows:

$$\mathrm{PE}^* = E_{\boldsymbol{\theta}}E_{\mathbf{X},\mathbf{Y}}[\mathbf{Y}-h(\mathbf{X},\boldsymbol{\theta})]^2. \tag{7}$$

Theorem 2. *For all θ, there are*

$$\mathrm{PE}^{**} \leq \overline{\rho}\mathrm{PE}^*. \tag{8}$$

In the equation, $\overline{\rho}$ is the weighted correlation coefficient of residual $\mathbf{Y}-h(\mathbf{X},\theta)$ and $\mathbf{Y}-h(\mathbf{X},\theta')$, and θ and θ' are independent of each other.

Theorem 2 provides the conditions for obtaining an accurate regression forest: low correlation between residuals and low-error decision tree. The random forest regression algorithm reduces the average error of the decision tree through the weighted correlation coefficient $\overline{\rho}$.

The steps of the random forest regression algorithm can be summarized as follows: let θ be a random parameter vector, and the corresponding decision tree is $T(\theta)$. Let $\mathbf{B}$ be the domain of $\mathbf{X}$; that is, $\mathbf{X}\colon \Omega\mapsto\mathbf{B}\subseteq\mathbf{R}^p$, where $p\in\mathbf{N}$ denotes the dimension of the independent variable. Each leaf node of the decision tree corresponds to a rectangular space. Remember the rectangular space of each leaf node as $\mathbf{R}_l\subseteq\mathbf{B}\,(l=1,2,\dots,L)$. For each $x\in\mathbf{B}$, if and only if one leaf node satisfies $\mathbf{x}\subseteq\mathbf{R}_l$, let the leaf node of the decision tree $T(\theta)$ be $l(\mathbf{x},\theta)$.

Step 1. Use the Bootstrap method to resample; randomly generate h training sets $\theta_1,\theta_2,\theta_3,\theta_4,\dots,\theta_{\mathbf{k}}$; and use each training set to generate the corresponding decision tree $\{T(\mathbf{x},\theta_1)\},\{T(\mathbf{x},\theta_2)\},\dots,\{T(\mathbf{x},\theta_{\mathbf{k}})\}$.

Step 2. Assuming that the feature has M dimensions, randomly extract m features from the M-dimensional features as the split feature set of the current node, and split the node in the best split method among the m features. Generally speaking, during the growth of the forest, the value of m remains unchanged.

Step 3. Each decision tree gets the maximum growth without pruning.

Step 4. For new data, the prediction of a single decision tree $T(\theta)$ can be obtained by averaging the observed values of the leaf node $\omega_{\mathbf{i}}(\mathbf{x})$. If an observation value X_i belongs to the leaf node $l(\mathbf{x})$ and is not 0, let the weight $\omega_{\mathbf{i}}(\mathbf{x})$ be

$$\boldsymbol{\omega}_i(x,\boldsymbol{\theta}) = \frac{1\{X_i\in\mathbf{R}_l(x,\boldsymbol{\theta})\}}{\#\left\{j\colon \mathbf{X}_j\in\mathbf{R}_l(\mathbf{x},\boldsymbol{\theta})\right\}}\,(i=1,2,3,\dots,n). \tag{9}$$

The sum of the weights is equal to 1.

Step 5. The prediction of a single decision tree is obtained by the weighted average of the observed value of the dependent variable $Y_i=(1,2,3,\dots k)$. The measured value of a single decision tree can be obtained using the following equation:

$$\widehat{\mu}(\mathbf{x}) = \sum_{i=1}^{n}\boldsymbol{\omega}_{\mathbf{i}}(\mathbf{x},\boldsymbol{\theta})\mathbf{Y}_i. \tag{10}$$

Step 6. Use (11) to obtain the weight of each observation by averaging the weight of the decision tree $\omega_i(\mathbf{x},\theta)$ $(t=1,2,\dots,k)$:

$$\boldsymbol{\omega}_{\mathbf{i}}(\mathbf{x}) = \frac{1}{k}\sum_{i=1}^{n}\boldsymbol{\omega}_i(\mathbf{x},\boldsymbol{\theta}_t)\mathbf{Y}. \tag{11}$$

Then, the predicted value of random forest regression can be recorded as follows:

$$\widehat{\mu}(\mathbf{x}) = \sum_{i=1}^{n}\boldsymbol{\omega}_{\mathbf{i}}(\mathbf{x},\boldsymbol{\theta})\mathbf{Y}_i. \tag{12}$$

The flowchart of the random forest algorithm is presented in Figure 2.

4. Prediction Platform Design

In view of the global spread of the COVID-19 epidemic in early 2020, we designed an epidemic prevention and control platform based on machine learning. The applications of this project include patient data analysis, early detection and warning of epidemic situations, rapid screening of suspected patients, and remote diagnosis and treatment. The patient data analysis application service must extract the patient's sign data for relevant detection, and the platform analyzes the data through the knowledge map to assist in the diagnosis in many aspects, significantly expediting virus diagnosis. The application service of early detection and early warning of epidemic situation needs to conduct regular and fixed-point investigation of disease data in various regions, conduct in-depth analysis by using knowledge map, and accurately grasp the warning of epidemic situation. The rapid screening application service for suspected patients collects data from urban infrastructure sensors and fixed-point sensors. The platform will match the data to the early symptoms of the virus, and people who are exposed to the virus and work in high-risk jobs will be given special attention to achieve the best resource allocation under the weight. Remote diagnosis and treatment application service for home isolation personnel, particularly suspected or close-contact personnel, can easily and quickly enter their information and achieve efficient and real-time remote diagnosis and control of the epidemic situation, and ordinary patients can be visited at home *via* remote diagnosis and treatment function, avoiding going to the hospital and

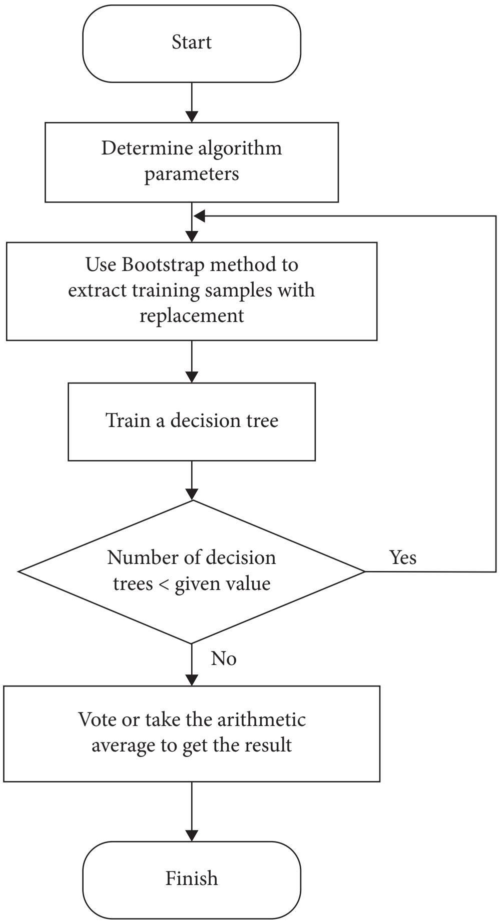
Start
Determine algorithm parameters
Use Bootstrap method to extract training samples with replacement
Train a decision tree
Number of decision trees < given value
Yes
No
Vote or take the arithmetic average to get the result
Finish

Figure 2: Random forest algorithm flowchart.

reducing the risk of infection. The epidemic prevention and control platform based on machine learning designed in this paper realizes the multiprocess integration of investigation, early warning, and diagnosis, and treatment after the outbreak of the epidemic achieves the unified analysis of various data and has the characteristics of accuracy, intelligence, and learnability. The platform architecture is presented in Figure 3.

During the overall integration of the system, the internal integration of each construction unit shall be done first, and then the integration of different construction units shall be carried out according to the interface definition of different system construction units and the order of strong to weak coupling or operation constraints. The interface definition method in the overall design shall be used to refer to the integration of various systems within each construction unit. The interfaces of various subsystems within the construction unit are clearly defined. On this basis, integration within the unit and between different units can follow one of two integration sequences, which can cross and coexist: the strength of coupling and the restriction of operation premise. Decomposition integration can be employed for integration between specific units. That is to decompose the integrated parts of this unit and other units and integrate them with relevant units in the form of a decomposition unit to reduce the complexity of integration and facilitate the positioning of problems. The overall integration test between the integration and relevant units must be conducted after each decomposition unit has completed the integration with specific integration objectives. Simultaneously, equipment and software products with good interconnection and interoperability must be selected. Moreover, attention must be paid during the development of application software to the interaction with other products to maintain consistency. In particular, the selection of a database requires a seamless connection with the heterogeneous database. The integrated system shall be convenient for expansion due to increased demand in the future.

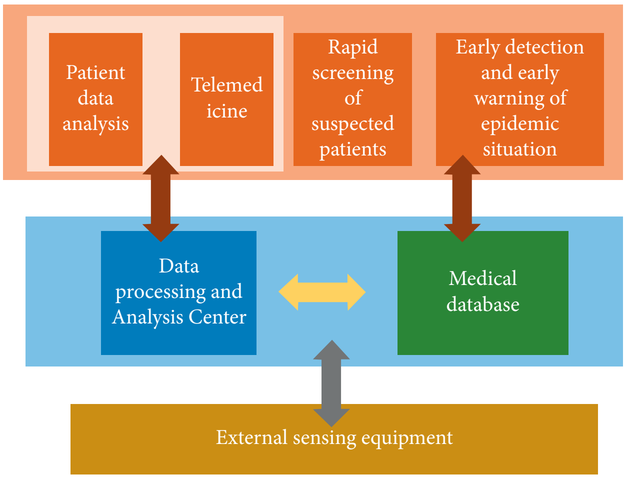
Patient data analysis
Telemed icine
Rapid screening of suspected patients
Early detection and early warning of epidemic situation
Data processing and Analysis Center
Medical database
External sensing equipment

Figure 3: Platform architecture.

4.1. Data Sources. The data sources involved in this project only include the hardware accessed by the platform, the data entered by the client, and the data of the original hospital systems (His, EMRs, PACS, and RIS), whereas the types of transmission data include user's physical sign data, basic information data, information interaction data, medical data, and publicity data of medical knowledge map. The medical knowledge map contains at least 5000 common disease data and 1000 virus-related data. This project involves structured, semistructured, and unstructured data. From the perspective of medical data storage, the overall data storage capacity of the medical industry is mainly 1–50 TB, and there are significant differences among medical institutions. For the time cycle of medical data, medical records are generally retained for a long time, and the requirements for online time are higher than those in other industries. The retention time of outpatient and emergency records shall not be less than 15 years, and the retention time of inpatient medical records shall be longer (about 30 years). The medical records of some celebrities will be kept indefinitely. Hundreds of image data must be stored and accessed during a patient's diagnosis activity. In general, clinical electronic medical record data use an XML file format that conforms to the standards, but the file format will continue to evolve. Medical data from two sources are stored in the medical database: one is the acquisition and input of the underlying hardware and the other is the medical data generated by the data analysis center. These data will be stored in a structured format, and if they are retrieved, they will be subject to permission access control. After granting access, the system will also collect visitor information to ensure the privacy, security, and traceability of medical data.

4.2. Data Access. The lower sensor is composed of medical sensing equipment, terminal equipment, information operation, and maintenance equipment using mobile medical perception technology. Wireless sensing technology, body area network technology, communication technology, terminal pass-through technology for telemedicine and positioning technology, monitoring network chip technology, and physiological signal acquisition and processing technology are examples of such technologies. Specifically, the lower sensor equipment is mainly composed of medical sensor equipment, terminal equipment, information operation, and maintenance equipment and is used for data collection and input of the prediction platform. In general, unstructured medical data is more serious and may have an impact on the storage quality of the database. The platform described in this article can handle hardware devices from a variety of ecological environments, and the output data is data structured using algorithms. Because of the diversity of unstructured mobile medical terminals, the prediction platform we designed supports a variety of access technologies and networking methods. By gathering multiple existing mobile medical terminals to form an enhanced virtual terminal, the prediction platform can be based on users. The environment where it is located automatically selects a suitable terminal device to access a specific wireless network and forms a multiterminal collaborative medical terminal system through virtual terminals formed by multiple terminals. The coordination of access, connection, transmission, and management of multiple heterogeneous network resources is referred to as multinetwork coordination. Because different medical systems use different heterogeneous networks for the transmission and use multiple wireless access technologies, it is important to overcome the limitations of a single network for multiple existing mobile medical terminals in order to achieve a more accurate and timely infectious disease prediction.

4.3. Multiplatform, Multisystem Data Normalization Processing, and Intelligent Analysis. Before inputting data to the random forest-based prediction platform designed in this research, the data needs to be preprocessed to facilitate the training and prediction of the neural network. Data normalization primarily refers to the distribution of experimental sample data into the intervals [0,1] or [−1,1] *via* multiplatform, multisystem, and heterogeneous health big data, so that the experimental sample data can be analyzed. The information is dimensionless. When collecting experimental data samples, this platform will generate unique sample data (singular sample data refers to the huge sample vector generated relative to other input sample data). Through this, the problem of gradient explosion and the subsequent decrease in the learning rate can be avoided. According to the data input requirements of the AI model and the neural network, an appropriate normalization method for health big data should be selected among the three commonly used normalization methods: min–max standardization, Z-score standardization method, and Z-score simplification. Moreover, the intelligent analysis of chronic disease data should be analyzed.

4.4. Data Security and Privacy Protection Methods. The security architecture of this prediction platform mainly includes application layer security, transport layer security, and perception layer security. The perception layer's security policies primarily include device authentication, data encryption, security coding, security protocols, and access control. Security strategies such as vulnerability scanning, active defense, security protocols, network filtering, and authorization management are mostly observed in the transport layer. The application layer mainly includes security policies and methods such as security auditing, intrusion detection, hot machine disaster recovery, virtual isolation, cloud antivirus, user permissions, and security management. The platform security architecture is presented in Figure 4.

5. RegressionValidation

The newly designed COVID-19 prediction platform in this paper includes the following five sections: data collection, data cleaning, machine learning training model, Bootstrap + Vue front-end framework, and Django back-end framework.

5.1. Data Collection andData Cleaning. Before model training, data collection and data cleaning are performed. The dataset contains 43 characteristic values, including monocyte percentage, monocyte count, lymphocyte count, platelet distribution width, and a list of label values. The dataset is randomly divided into an 80% training set and a 20% test set using the Numpy matrix operation library and the pandas library based on Numpy for data processing. The process is shown in Figure 5.

5.2. Machine Learning Training Models. The training models are created using the logistic regression algorithm, SVM, and random forest. Both the training and test sets have parameter passing settings [15, 16]. The primary goal of machine learning is to obtain a prediction model by mining the inherent patterns in historical training data and then applying the model to similar data situations [17, 18]. The general workflow diagram is presented in Figure 6.

In Section 6, the model training mechanisms and prediction accuracies are simulated and compared.

5.3. Bootstrap + Vue Front-End Framework. After model training is complete, model visualizations are compared. The sign-in and registration pages are shown in Figure 7. For

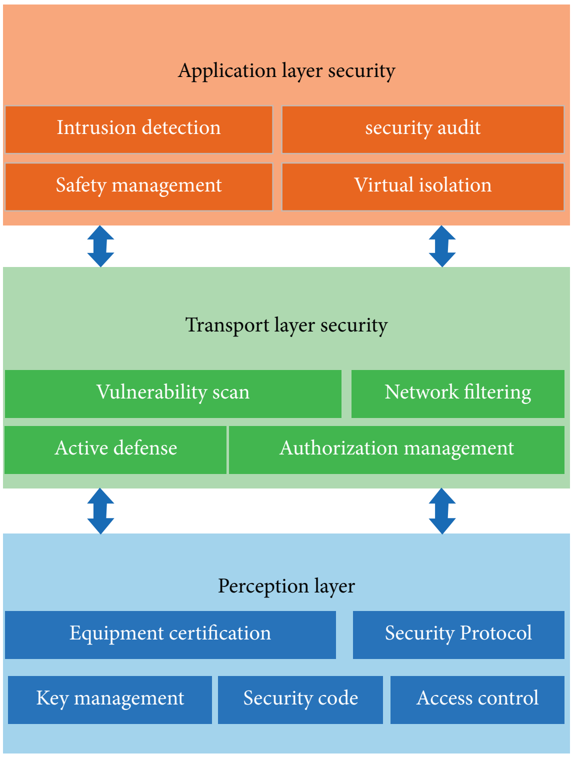

Figure 4: Platform security architecture.

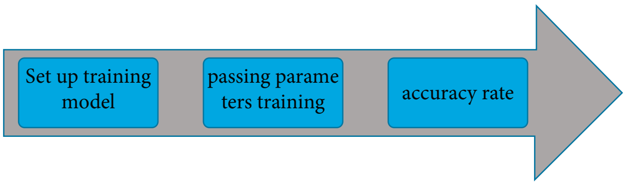

Figure 5: Matrix operation based on Numpy.

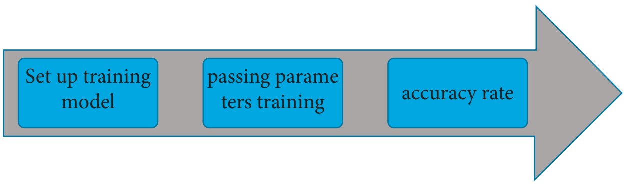

Figure 6: The training procedure.

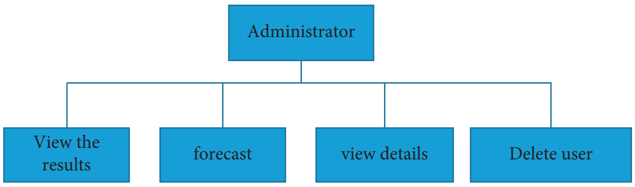

Figure 7: The training procedure.

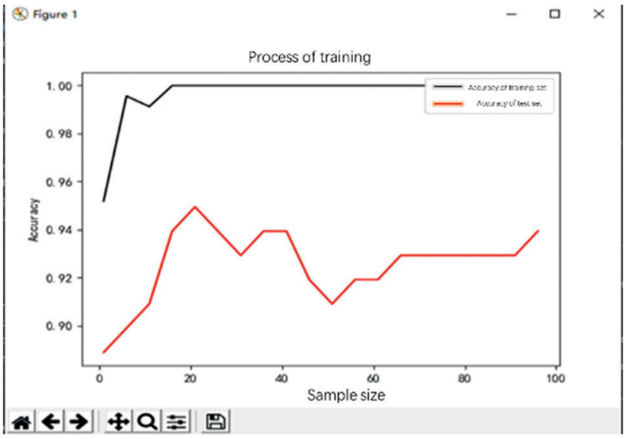

Figure 8: Decision tree number.

legality verification, e-mail and password must be entered and sent back to the server. The registration information from the registration page is sent to the background *via* a POST request and saved in the database using the Django ORM model. The default registration is a regular user who can only perform the detection function. After entering the system, a relevant interface will appear, which requires nonrepeated patient numbers and presents the prediction information of all patients. After users click the detection button in the upper left corner, the detection model box will appear. Users must enter patient information, such as whether the patient has a fever or COVID-19, as well as routine blood test results. The data will be sent to the background for prediction processing, and the results will be returned.

5.4. Django Back-End Framework. Django background is primarily used for providing a front-end request interface and returning values required by the front-end template, as well as managing the user permission in the meantime.

Data, including user models, user details, and front-end homepage display information, are stored in the database through the ORM model and returned to the foreground (Figure 7).

6. ModelPrediction Simulation

The simulation in this paper is based on the epidemic dataset containing more than 40 kinds of characteristic data, including lymphocyte percentages, classifying whether the disease is classified, setting up the five middle training models of random forest and neural future, comparing the accuracy and other characteristics, and concluding.

6.1. Simulation Analysis. Once the models are complete, the parameters required by each algorithm, such as tree number and maximum depth in the model, need to be tuned.

Figures 8 and 9present the influences of the aforementioned parameters in the random forest training model on the test set accuracy, respectively. Two parameters are tuned simultaneously in the actual parameter tuning process to select the best accuracy of a test set. Simultaneously, the same parameter tuning process is applied to the other training models, which tunes the parameters with the greatest influence on each algorithm for each optimal

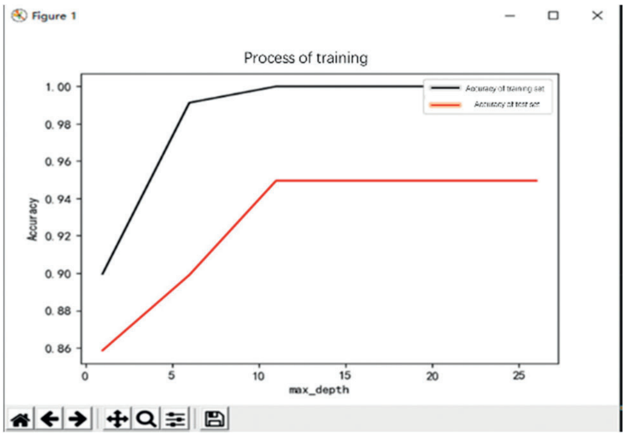

Figure 9: Maximum depth.

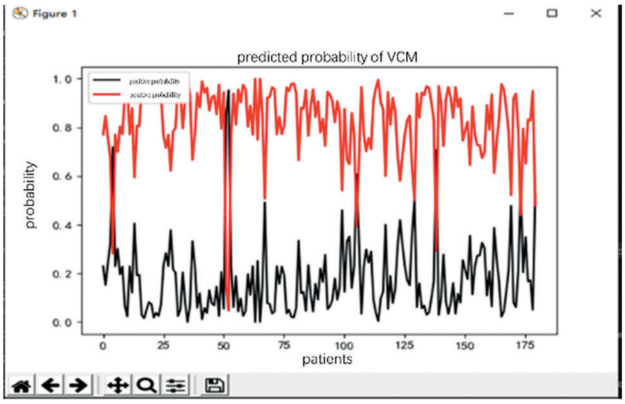

Figure 10: SVM training model.

accuracy selection and saves the parameters with the best values as the local models. Finally, the simulation results that are predicted from the datasets that use the real patient information are shown below.

As illustrated in Figures 10–12, random forest is selected as the best method for prediction analysis in this scenario when compared with the other methods, each of which has advantages and disadvantages. In Figures 10–12, the red lines indicate the positive probability, and the black lines represent the negative probability.

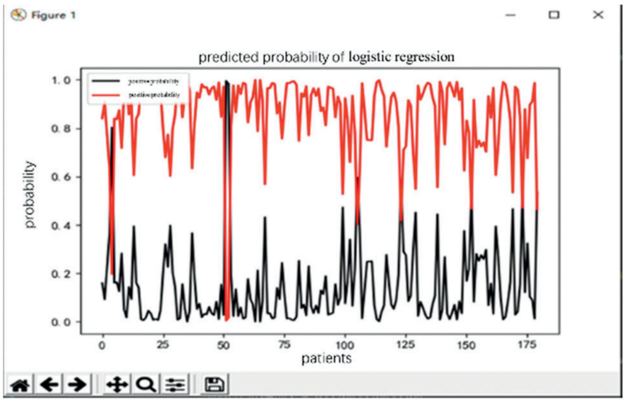

Figure 11: Logistic regression model.

6.2. The Simulation Principle of Random Forest. This algorithm uses random sampling with the replacement method to select the training set and builds the classifiers accordingly. In addition, multiple decision trees are established and merged for more accurate and stable predictions. Finally, the best classification results are selected.

By voting, the principle of the random forest algorithm is presented in Figure 13.

The foundation of random forest is Bootstrap. That is, many new samples of the same size and usability are generated from a sample, and similar samples are generated again from those that have already been created. Bootstrap is also known as a self-help method as it does not use any other sample data [19]. When the sample size is small, this method is considered useful. If the traditional method is used for verifications and segmentations, the sample size will be even smaller, resulting in a larger deviation and a nonoptimal solution[20]. The self-service method not only fails to reduce the training sample size but also leaves.

A validation set: the random forest algorithm is the integration of bagging and decision trees. After multiple samplings, partial samples cannot be extracted from the training set. These unsampled data are called out of the bag (OOB). OOB is not added into the training set by the model

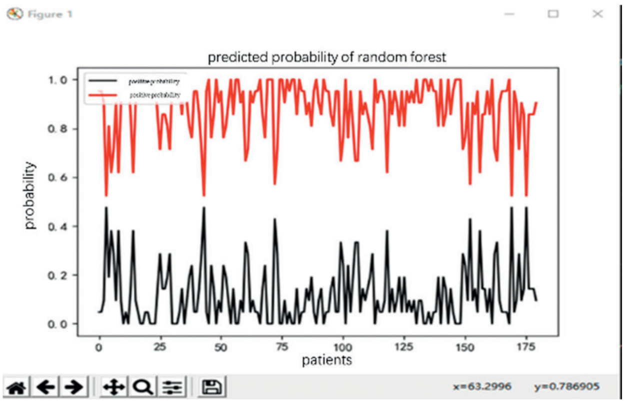

Figure 12: Random forest training model.

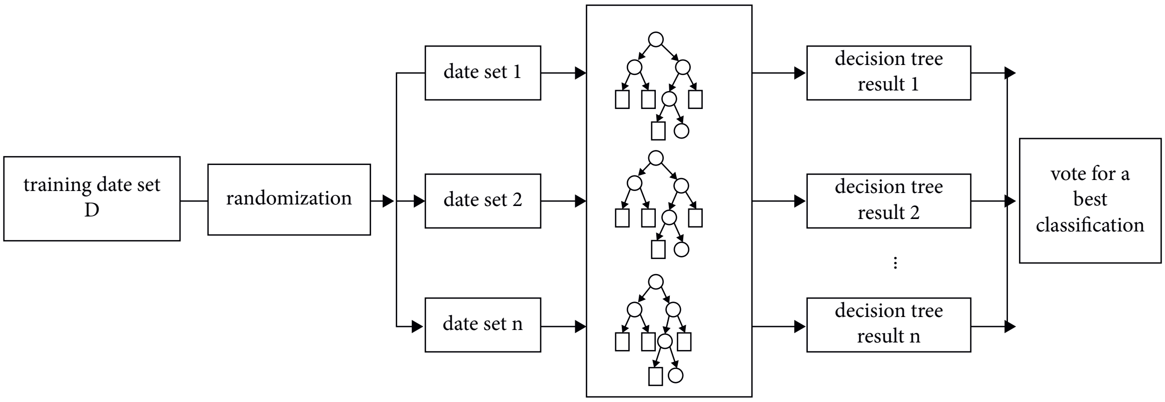

Figure 13: Random forest algorithm principle diagram.

for fitting, which makes it applicable for the detection of the model generalization ability [21, 22].

7. Conclusions

In this paper, the COVID-19 prediction algorithms based on artificial intelligence were compared. Based on considerations of various characteristic constraints and prediction result accuracies, a prediction platform was established. Through simulation, it was discovered that random forest has significant advantages in epidemic prediction over logistic regression and the support vector machine. It would perform admirably when applied to the medical platform designed in this paper. Simultaneously, Singh proposed using the unmanned aerial vehicle [23] based on blockchain to achieve contactless transmission in the COVID-19 environment [24, 25]. Similar application scenarios such as [26] will be the next development direction of the platform and strived to support more application development in the epidemic environment based on prediction analysis [27].

Data Availability

The patient data used to support the findings of this study are restricted by The First Affiliated Hospital of Nanjing Medical University in order to protect patient privacy. The data are available from The First Affiliated Hospital of Nanjing Medical University for researchers who meet the criteria for access to confidential data.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The work was supported by the National Key R&D Program of China (2018YFC1314900) and the Key R&D Program of Jiangsu (BE2020721).

References

[1] G. Smith and C. Cheeseman, “A mathematical model for the control of diseases in wildlife populations: culling,vaccination and fertility control,” *Ecological Modelling*, vol. 150, no. 1-2, pp. 45–53, 2002.

[2] A. Assiri, “Anomaly classification using genetic algorithm-based random forest model for network attack detection,” *Computers, Materials & Continua*, vol. 66, no. 1, pp. 767–778, 2020.

[3] A. Js, A. Mf, B. Bc, and C. Bl, “Noncontact and nondestructive evaluation of heat-treated bearing rings using pulsed eddy current testing,” *Journal of Magnetism and Magnetic Materials*, vol. 511, 2020.

[4] A. Althobaiti, A. Jindal, and A. K. Marnerides, “Scada-agnostic power modelling for distributed renewable energy sources,” in *Proceedings of the 2020 IEEE 21st International Symposium on ”A World of Wireless, Mobile and Multimedia Networks*, Cork, Island, June 2020.

[5] J. Lin, X. He, S. Lu, D. Liu, and P. He, “Investigating the Influence of Three-Dimensional Building Configuration on Urban Pluvial Flood-Ing Using Random forest Algorithm - Sciencedirect,” *EnvironmentalResearch*, vol. 196, 2020.

[6] B. Silke, J. Kellett, T. Rooney, K. Bennett, and D. O’Riordan, “An improved medical admissions risk system using multivariable fractional polynomial logistic regression modelling,” *QJM: International Journal of Medicine*, vol. 103, no. 1, pp. 23–32, 2009.

[7] K. Korpela, M Renko, N Paalanne et al., “Microbiome of the first stool after birth and infantile colic,” *Pediatric Research*, vol. 88, no. 1, pp. 776–783, 2020.

[8] Wikipedia, “Decision tree learning,” https://en.wikipedia.org/wiki/Decisiontreelearn/ ”.

[9] A. Jindal, A. Dua, K. Kaur, M. Singh, N. Kumar, and S. Mishra, “Decision tree and svm-based data analytics for theft detection in smart grid,” *IEEE Transactions on Industrial Informatics*, vol. 12, no. 3, pp. 1005–1016.

[10] T. Xiao, T. Zhang, and N. Zhang, “Green production cycle mining of mass production based on random forest algorithm,” *International Journal of Product Development*, vol. 24, 2020.

[11] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[12] G. W. Cha, H. J. Moon, Y. M. Kim, W. H. Hong, J. H. Ha, W. J. Park, Y. C. Kim, Development of a prediction model for demolition waste generationusing a random forest algorithm based on small datasets,” *InternaTional Journal of Environmental Research and Public Health*, vol. 17, no. 19, 2020.

[13] M. Fang, F. Liu, L. Huang, L. Wu, L. Guo, and Y. Wan, “A urine metabonomics study of rat bladder cancer by combining gas chromatography-mass spectrometry with random forest algorithm,” *International Journal of Analytical Chemistry*, vol. 2020, no. 9, 9 pages, Article ID 8839215, 2020.

[14] I. Hernández and W. Yang, S. Mohan and N. Jowett, Label-free histo- morphometry of peripheral nerve by stimulated Raman spectroscopy,” *Muscle & Nerve*, vol. 62, no. 1, 2020.

[15] H. Cao, A. Xiao, Y. Hu, P. Zhang, S. Wu, and L. Yang, “On virtual resource allocation of heterogeneous networks in virtualization environment: a service oriented perspective,” *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 4, pp. 2468–2480, 1 Oct.-Dec. 2020.

[16] H. Cao, L. Yang, and H. Zhu, “Novel node-ranking approach and multiple topology attributes-based embedding algorithm for single-domain virtual network embedding,” *IEEE Internet of Things Journal*, vol. 5, no. 1, pp. 108–120, 2018.

[17] H. Cao, J. Du, H. Zhao et al., “Resource-ability assisted service function chain embedding and scheduling for 6G networks with virtualization,” *IEEE Transactions on Vehicular Technology*, vol. 70, no. 4, pp. 3846–3859, 2021.

[18] H. Cao, S. Wu, G. S. Aujla, Q. Wang, L. Yang, and H. Zhu, “Dynamic embedding and quality of service-driven adjustment for cloud networks,” *IEEE Transactions on Industrial Informatics*, vol. 16, no. 2, pp. 1406–1416, 2020.

[19] T. A. Rather, S. Kumar, and J. A. Khan, “Multi-scale habitat modelling and predicting change in the distribution of tiger and leopard using random forest algorithm,” *Scientific Reports*, vol. 10, no. 1, Article ID 11473, 2020.

[20] M. Zheng, W. Tang, A. Ogundiran, and J. Yang, “Spatial simulation modeling of settlement distribution driven by random forest: consideration of landscape visibility,” *Sustainability*, vol. 12, no. 11, 2020.

[21] P. . Calhoun, R. A. Levine, and J. Fan, “Repeated measures random forests (rmrf): identifying factors associated with nocturnal hypo- glycemia,” *Biometrics*, vol. 77, no. 1, 2020.

[22] F. Fontove and G. D. Rio, “Residue cluster classes: a unified protein representation for efficient structural and functional classification,” *Entropy*, vol. 22, no. 4, 2020.

[23] H. Cao, Y. Hu, and L. Yang, “Towards intelligent virtual resource allocation in UAVs-assisted 5G networks,” *Computer Networks*, vol. 185, Article ID 107660, 2021.

[24] L. Liu, J. Feng, Q. Pei et al., “Blockchain-Enabled secure data sharing scheme in mobile-edge computing: an asynchronous advantage actor-critic learning approach,” *IEEE Internet of Things Journal*, vol. 8, no. 4, pp. 2342–2353, 2021.

[25] J. Feng, F. R. Yu, Q. Pei, X. Chu, J. Du, and L. Zhu, “Cooperative computation offloading and resource allocation for blockchain- enabled mobile-edge computing: a deep reinforcement learning approach,” *IEEE Internet of Things Journal*, vol. 7, no. 7, pp. 6214–6228, 2019.

[26] M. Singh, G. S. Aujla, R. S. Bali, S. Vashisht, and A. Jindal, “Blockchain-enabled secure communication for drone delivery: a case study in covid-like scenarios,” in *Proceedings of the 2nd ACM MobiCom Workshop on Drone Assisted Wireless Communications for 5G and Beyond*, London, UK, September 2020.

[27] H. Cao, H. Zhu, and L. Yang, “Collaborative attributes and resources for single-stage virtual network mapping in network virtualization,” *Journal of Communications and Networks*, vol. 22, no. 1, pp. 61–71, 2020.