A Hybrid Feature Optimized Framework for Enhanced COVID-19 Detection with IoT Sensor Data

P. Santosh Kumar Patra[1*], Biswajit Tripathy[2]

[1]Research Scholar, Department of Computer Science and Engineering, Biju Patnaik University of Technology, Rourkela, Odisha, 769015, India

[2]Professor, Master of Computer Applications, Einstein College of Computer Application and Management, Khurda, Odisha, 752060, India

E-mail: psantoshkumarpatra1@gmail.com[1], biswajit69@gmail.com[2]

*Corresponding author: P. Santosh Kumar Patra

Abstract

The COVID-19 pandemic has profoundly affected global public health, requiring the creation of precise and effective illness detection tools. The expansion of the healthcare sector has been significantly enhanced by the Internet of Things (IoT), which supports various applications such as telemedicine and direct consultations. COVID-19 can be readily identified by employing artificial intelligence algorithms on users' IoT data. Nevertheless, conventional artificial intelligence algorithms were inadequate in extracting and selecting features from the dataset. This study employed the machine learning optimized COVID-19 classification model to identify SC2, other, and no virus categories from IoT data. The dataset undergoes pre-processing to eliminate noise, address missing values, and remove redundant features, so ensuring data quality. The Random Forest Infused Black Widow Optimization (RFI-BWO) technique is subsequently utilized to extract pertinent features from the pre-processed dataset, harnessing the efficacy of random forest and the optimization potential of the Black Widow Optimization algorithm. Subsequent to feature extraction, the Linear Logistic Regression-based Genetic Optimization (LLRGO) technique is employed for feature selection, identifying the most pertinent features that significantly contribute to COVID-19 detection. LLRGO seeks to improve classification performance and diminish computational complexity through the iterative selection and evaluation of feature subsets. The chosen characteristics are input into the Custom Convolutional Neural Network (CCNN), which learns intricate patterns and correlations among the specified features, facilitating precise classification of COVID-19 categories. The simulation conducted on a San Francisco COVID-19 dataset demonstrates that the suggested proposed model attained an accuracy of 97.87%, precision of 98.41%, and F1-score of 96.96%, surpassing traditional approaches.

Keywords: Internet of Things, COVID-19 Classification, Random Forest Infused Black Widow Optimization, Linear Logistic Regression-based Genetic Optimization, Custom Convolution Neural Network

1. Introduction

COVID-19, a pandemic instigated by the newly identified coronavirus, originated in late 2019 in Wuhan, China. Data on COVID-19 cases and fatalities reported by the World Health Organization (WHO) across various nations. The cumulative total of cases refers to the number of confirmed COVID-19 cases recorded since the onset of the pandemic in each country. The cumulative sum of COVID-19 cases per 100,000 population columns reflects the number of cases per 100,000 individuals in each country [2]. This facilitates the comparison of numerous cases across countries with varying populations. The newly reported instances in the last 7 days denote the quantity of freshly documented COVID-19 cases within the preceding week. The number of newly reported COVID-19 cases in the last 7 days per 100,000 population indicates the incidence of new cases per 100,000 individuals over the preceding week [3]. The number of new COVID-19 cases reported in the last 24 hours indicates the recent cases. The cumulative sum of deaths represents the aggregate number of fatalities attributed to COVID-19 since the beginning of the epidemic in each respective country. The death column reflects the cumulative sum of COVID-19 fatalities per 100,000 individuals in each country. This facilitates a standardized comparison of mortality rates across nations with varying populations. The newly reported deaths in the past week represent the COVID-19 fatalities documented throughout that period. The number of newly reported COVID-19 deaths per 100,000 population in the last 7 days. The observations underscore the necessity for early diagnosis of COVID-19 across several nations for comparative investigation of the pandemic [4]. The healthcare industry has also witnessed the emergence of specialized facilities, such as intensive care units, to cater to the increasing demand for critical care. These facilities, found not only in hospitals but also in various healthcare settings, provide essential support for patients in need of intensive medical attention [5]. Nursing programs focusing on healthcare equip students with up-to-date knowledge and skills that can be applied across diverse healthcare settings, ensuring they are prepared to contribute effectively to the evolving healthcare landscape. The growth of the healthcare industry has been further facilitated by the IoT [6], which enables a wide range of applications, including telemedicine and direct consultation, which enhance healthcare accessibility and efficiency. These technologies allow for remote monitoring of patients, virtual consultations with healthcare professionals, and the seamless integration of medical devices and systems, all of which contribute to improved healthcare outcomes.

The COVID-19 pandemic has underscored the pressing necessity for expedited and more effective measures to counteract the swift transmission of the virus and mitigate the burden on healthcare institutions. A possible approach to tackle these difficulties is the utilization of machine learning, the Internet of Things, and artificial intelligence [7]. Machine learning, a subset of artificial intelligence, has the capacity to reveal new patterns and insights inside extensive datasets. Machine learning employs computer algorithms to discern links and derive important insights from various datasets, irrespective of their size or distinctiveness. Considering the vast volume of data amassed daily regarding the COVID-19 epidemic globally, machine learning methodologies can be essential in mining and evaluating this information to derive meaningful and pertinent insights. This can significantly improve decision-making processes and assist in formulating effective methods to combat the virus [9]. Machine learning is becoming prevalent in the medical area, with applications across diverse industries and situations. For instance, it can assist in averting contamination, assessing treatment efficacy, accessing essential data, and further functions. Its adaptability and capacity for progress render it a crucial instrument in combating COVID-19 and addressing future healthcare concerns. The COVID-19 pandemic has distinct problems in comparison to prior viral outbreaks. The intricate dynamics of populations, the varied measures employed by governments, and the proclamation of states of emergency have generated ambiguities in the virus's spread. These characteristics complicate reliance on current prediction models [10], underscoring the necessity for ongoing modification and enhancement of these models to precisely reflect the pandemic's dynamics.

The researchers Arwolo et al. [11] devised a solution for COVID-19 outbreaks based on machine learning and the IoT. This work utilized both artificial bee colony (ABC) optimization for feature extraction and support vector machine (SVM) for classification. Furthermore, they developed the L-SVM-ABC and Q-SVM-ABC models by combining two individual models. However, one drawback of their approach was the need for extensive computational resources and time-consuming optimization techniques. Ibrahim et al. [12] proposed multi-region machine learning-based ensemble methodologies to provide accurate predictions. They included models such as adaptive neuro-fuzzy inference systems (ANFIS), artificial neural networks (ANN), SVM, multinomial naïve bayes (MNB) and traditional multiple linear regression models (MLR). However, a limitation of their approach was the complexity of integrating and fine-tuning multiple models, which required careful parameter selection and training. Zong et al. [13] developed a model for the distribution of COVID-19 resources based on the concept of reinforcement learning. They created an agent-based epidemic environment to investigate transmission dynamics through IoT. One drawback of their approach was the reliance on accurate and up-to-date data for training the reinforcement learning model, which could be challenging during rapidly evolving outbreaks. Ahmed et al. [14] created an IoT-enabled smart healthcare system for the automated detection and classification of infectious disorders (pneumonia, COVID-19) in chest X-ray images. Their system employed two unique deep learning architectures and a method that utilized multiple layers of feature fusion and selection. However, a limitation was the requirement for a large and diverse dataset for training deep learning models, which does not always be readily available. Almagrabi et al. [15] presented an RL-based crowd-to-machine (RLC2M) architecture for mH-IoT, addressing challenges related to processing healthcare information. They utilized crowdsourcing in conjunction with a reinforcement learning model known as Q-learning. One drawback of their framework was the potential for bias or inconsistency in the data obtained from crowdsourcing, which could impact the accuracy and reliability of the learned model.

3. Proposed Methodology

The proposed model is implemented for detecting COVID-19 cases from IoT data. Figure 1 shows the proposed block diagram with multiple stages. The dataset preprocessing involves handling missing values, removing any noisy data, and eliminating redundant features. By ensuring data quality, the subsequent steps can be more effective in extracting meaningful patterns. Then, RFI-BWO algorithm is used to extract relevant features from the pre-processed dataset. RFI-BWO combines the strengths of two techniques: random forest, which is an ensemble learning method capable of capturing complex relationships in data, and the Black Widow Optimization algorithm, which is an optimization algorithm inspired by the hunting behaviour of black widow spiders.

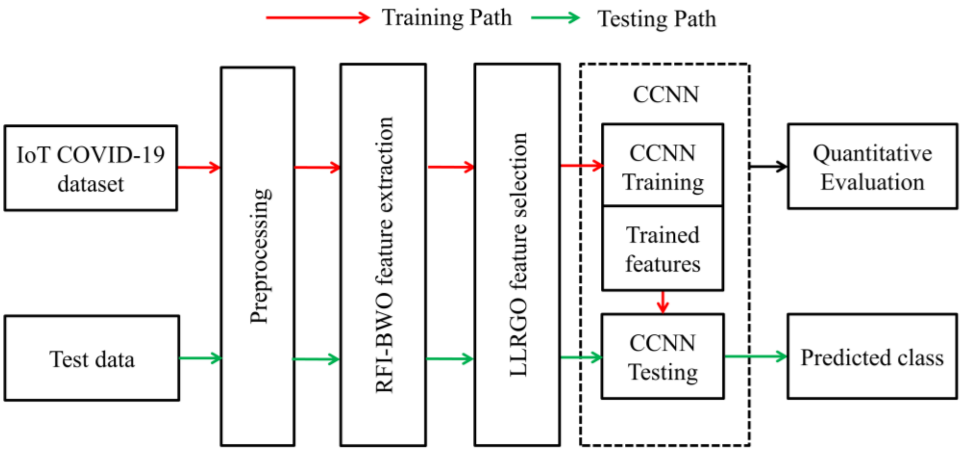


Figure 1. Proposed block diagram.

This feature extraction process aims to identify the most informative features for COVID-19 classification. The LLRGO algorithm is employed for feature selection. LLRGO uses linear logistic regression, a type of classification algorithm, as the fitness function within a genetic optimization framework. By iteratively selecting and evaluating subsets of features, LLRGO aims to identify the most relevant features that contribute significantly to COVID-19 detection. This step helps to reduce the computational complexity and improve the classification performance. Finally, the selected features are fed into a CCNN, which is a type of deep learning model particularly effective in analyzing textual data. In this case, the CCNN is designed specifically for the selected features. The CCNN learns complex patterns and relationships within the feature set, enabling accurate classification of COVID-19 cases.

3.1 Data preprocessing

The data preprocessing is a preliminary procedure but an essential one in proposed method. This involves removing any anomalous data, redundant data, and outliers. The method of scaling features that is known as min–max normalization was used at the pre-processing stage of the analysis. The data that was obtained are subjected to a linear transformation as a result of using this method. This method adjusts the values of the data such that they fall somewhere between 0 and 1, while also preserving the connection between the newly collected data points. The use of this constrained range has been done with the expectation that it would culminate in extremely modest standard deviations, which will obliterate the impact of any outliers. The min–max normalization formula is shown in the following.

$$y' = \frac{y - y_{min}}{y_{max} - y_{min}} \qquad (1)$$

Here, value y' represents the resultant value, the values y_{min} and y_{max} correspond to the lowest and maximum values of the dataset that is being used.

3.2 RFI-BWO feature extraction

The behaviour of the RFI-BWO was the inspiration for the black spider's nature, which is a bio-inspired combinatorial optimization method. Figure 2 shows the proposed RFI-BWO feature extraction algorithm flowchart, and Table 1 shows the algorithm steps of RFI-BWO. Spiders are a widespread kind of arthropod that was identified by its eight legs and, in certain circumstances, their poisonous fangs. Arachnids in general are nocturnal creatures, and black widow spiders are no exception; they construct their webs during the dark hours of the night. It is common for females to spend most of their life in a single location, and while they are there, they emit hormones that attract men. The female black widow spider is known for her cannibalistic tendencies, as she will eat the male spider either during or after the fertilization of her eggs. This is how the spider got its name. Nevertheless, cannibalism has been seen in these spiders on several occasions, as their offspring feed on each other and, in some situations, even swallow their own mother. As a result, in this approach of optimization, only the most proficient spiders are allowed to continue, which enables the resolution of difficult issues by modelling this behaviour through evolution rules.

Table 2. RFI-BWO feature extraction algorithm.

Input: Preprocessed Dataset

Output: RFI-BWO features

Step 1: Initialize the algorithm: set the population size. Define the maximum number of iterations.

Step 2: Initialize the population

Step 3: Evaluate the fitness of each widow in the population using the random forest model.

Step 4: Identify the widow with the highest fitness value as the global best solution.

Step 5: Perform the optimization process: Enter the main optimization loop, which continues until the maximum number of iterations is reached or the convergence criteria are satisfied.

> **Step 5.1:** Selection and reproduction: Select a subset of widows for reproduction based on their fitness values.
>
> **Step 5.2:** Apply a selection mechanism, such as tournament selection or roulette wheel selection, to choose the parent widows.
>
> **Step 5.3:** Crossover: Perform crossover between the selected parent widows to generate offspring.
>
> **Step 5.4:** Apply a crossover operator, such as single-point crossover or uniform crossover, to create new solutions.
>
> **Step 5.5:** Maintain the diversity in the population by preserving some parent solutions.
>
> **Step 5.6:** Mutation: Apply a mutation operator to introduce small random changes in the offspring solutions. The mutation operator can be applied to selected attributes or parameters of the widows.
>
> **Step 5.7:** Evaluation and replacement: Evaluate the fitness of the offspring widows using the random forest model.
>
> **Step 5.8:** Replace the least fit widows in the population with the newly generated offspring.
>
> **Step 5.9:** Update the best solution: If any offspring widow has a higher fitness value than the current global best, update the global best solution.
>
> **Step 5.10:** Convergence check: Check if the convergence criteria are met.
>
> **Step 5.11:** If the criteria are satisfied (e.g., no significant improvement in the global best solution for a certain number of iterations), terminate the optimization process.

Step 6: Output the results: Once the optimization process is completed, output the best solution found, along with its fitness value and features are extracted.

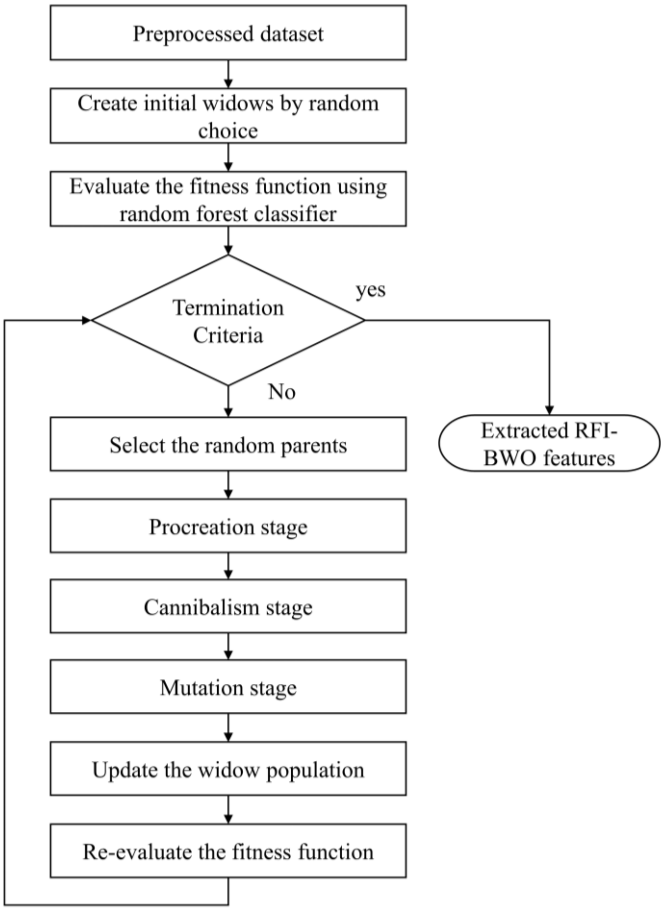

Preprocessed dataset
Create initial widows by random choice
Evaluate the fitness function using random forest classifier
Termination Criteria
yes
No
Extracted RFI-BWO features
Select the random parents
Procreation stage
Cannibalism stage
Mutation stage
Update the widow population
Re-evaluate the fitness function

Figure 2. RFI-BWO feature extraction flowchart.

3.3 LLRGO feature selection

A search-based optimization strategy called the genetic algorithm was developed using genetics and natural selection as its primary sources of inspiration. In the process of feature selection, a LLRGO was used to pick the subset of characteristics that are most relevant to a certain machine-learning job. The use of a LLRGO for feature selection is beneficial since it can determine the ideal subset of features, which ultimately results in enhanced performance and less computing cost. Table 2 shows the algorithm steps of LLRGO.

Table 2. RFI-BWO feature extraction algorithm.

Input: RFI-BWO features **Output:** LLRGO selected features
Step 1: Initialize the population: Set the population size, which determines the number of individuals (feature subsets) in each generation. **Step 2:** Generate an initial population of feature subsets. Everyone represents a potential solution (set of features) and is encoded as a binary string, where each bit corresponds to the presence or absence of a feature. **Step 3:** Evaluation: Evaluate the fitness of everyone in the population by training a logistic regression model on the selected features and measuring accuracy through cross-validation. **Step 4:** Assign a fitness value to everyone based on its performance. **Step 5:** Selection: Choose a subgroup of individuals from the population for reproduction depending on their fitness values. Individuals with more physical prowess are more likely to be chosen.

Step 6: Apply a selection mechanism, such as tournament selection or roulette wheel selection, to choose the parent individuals for the next generation.

Step 7: Reproduction: Perform crossover between the selected parent individuals to generate offspring. Crossover entails the transfer of genetic information (bits) between parents in order to generate novel solutions.

Step 8: Apply a crossover operator, such as single-point crossover or uniform crossover, to create new feature subsets. The crossover probability determines the likelihood of performing crossover.

Step 9: Mutation: Employ a mutation operator to induce minor random alterations in the offspring solutions. The mutation operator can be applied to selected attributes (bits) of the feature subsets.

Step 10: Mutation helps to introduce diversity and explore different parts of the solution space. The mutation probability determines the likelihood of performing mutation.

Step 11: Evaluation and replacement: Evaluate the fitness of the offspring individuals (feature subsets) using the logistic regression model and the chosen performance metric.

Step 12: Substitute the individuals with the lowest level of fitness in the existing population with the newly produced offspring individuals.

Step 13: Termination condition: Check for termination conditions. If the maximum number of generations (iterations) has been reached, or there is no significant improvement in the fitness value over a specified number of iterations.

Step 14: If the termination condition is met, stop the optimization process. Otherwise, go back to Step 3 (Selection) and continue to the next generation.

Step 15: Output the results: Once the optimization process is completed, select the best individual (feature subset) based on its fitness value.

3.4 CCNN classification

CCNN is a technique of deep learning that was used to a variety of tasks, including object identification, data classification, and other computer vision-related tasks. The CCNN employs a multi-layer perceptron of various convolution layers. The convolution layer is the core component of the CCNN design and is responsible for the feature extraction process that is carried out on the input data. The convolutional layer makes use of several parameters, the most important of which are the padding, kernel size, filter, and stride. The MaxPooling layer will down sample the data in order to lower the amount of computation required as well as the network parameters. The process of pooling is a kind of down-sampling that involves reducing the size of each feature map in order to lessen the likelihood of overfitting. The maximum pooling method and the average pooling method are the two primary classifications of pooling operations. The Fully Connected Layer (FC) and the SoftMax function have the capacity to classify COVID19 with probability values ranging from 0 to 1.

4. Results and discussion

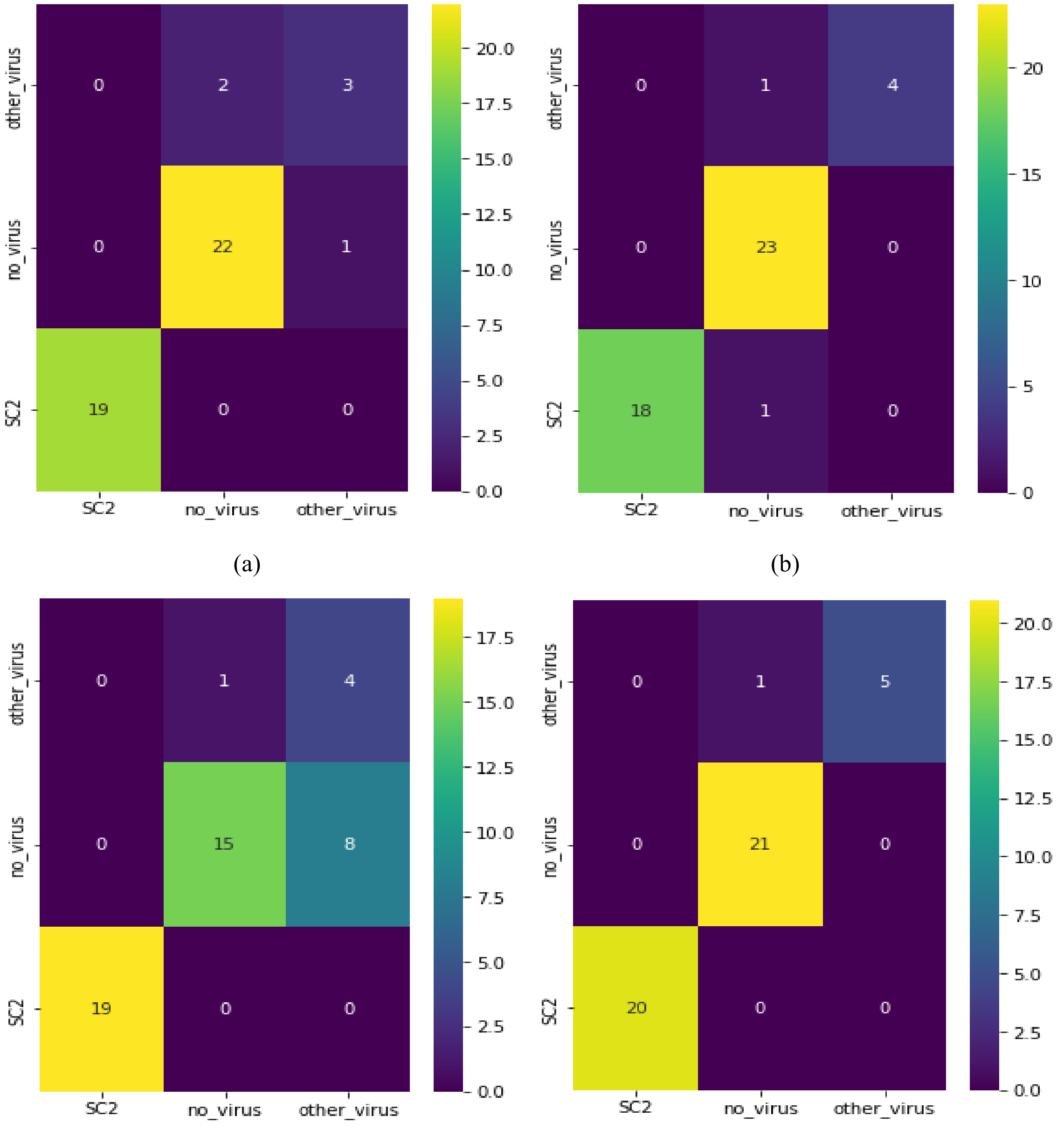
This section provides a comprehensive examination of the simulation results obtained from the proposed model. Further, the performance of proposed model is compared with state of art methods using same dataset. In addition, various performance measures are evaluated and compared.

4.1 Dataset

The IoT-San Francisco COVID-19 dataset contains the various columns such as CZB_ID, sequencing_batch, gender, age, SC2_PCR, SC2_rpm, idseq_sample_name, and viral_status. Each row represents a sample of patient data. Each row represents a different sample, and the information provided includes whether the sample tested positive or negative for SARS-CoV-2, the abundance of viral reads, and the presence of other viruses. The dataset is collected from 234 different persons with all column's information. For an instance, the no virus class contain 100 records, SC2 class contain 93 records, and other_virus class contain 41 records.

4.3 Performance evaluation

In Figure 6, the confusion matrices of different methods, namely L-SVM-ABC [11], Q-SVM-ABC [11], MNB [12], and the proposed model, are depicted. A confusion matrix is a tabular representation that provides a concise summary of a classification model's performance. It displays the number of correct positive predictions, correct negative predictions, incorrect positive predictions, and incorrect negative predictions. The proposed model had the best accuracy score, indicating a greater percentage of properly identified examples compared to other methods. The L-SVM-ABC [11] method, despite achieving relatively low accuracy, have limitations such as a higher computational complexity and longer training time. While the MNB [12] method achieves a reasonable level of accuracy, it has limitations in handling complex relationships and interactions between features in the data. This is because the MNB [12] algorithm assumes independence among features, which not hold true in certain scenarios, potentially leading to suboptimal performance when dealing with intricately correlated data.



(a)

(b)

(c)

(d)

Figure 6. Confusion matrices of various methods. (a) L-SVM-ABC [11]. (b) Q-SVM-ABC [11]. (c) MNB [12]. (d) Proposed model.

Table 4 compares the performance of proposed model with existing approaches. Here, the proposed model resulted in improved performance as compared to L-SVM-ABC [11], Q-SVM-ABC [11], and MNB [12]. The proposed method shows an improvement of 19.13% in accuracy, 9.41% in precision, 27.18% in F1-score compared to L-SVM-ABC [11]. Compared to Q-SVM-ABC [11], the proposed method demonstrates a significant improvement of 2.13% in accuracy, 1.31% in precision, 0.95% in F1-score. The proposed method surpasses MNB [12] with a substantial improvement of 23.41% in accuracy, 26.19% in precision, and 27.52% in F1-score.

Table 4. Performance comparison of proposed model with existing approaches.

Method	Accuracy	Precision	F1-Score
L-SVM-ABC [11]	78.723	89.245	69.78
Q-SVM-ABC [11]	95.74	97.10	95.89
MNB [12]	74.46	72.22	69.44
Proposed model	97.87	98.41	96.96

5. Conclusion

The project effectively deployed the proposed model for the identification of SC2, other, and no virus categories from IoT data. The dataset was initially pre-processed to assure data quality by minimizing noise, addressing missing values, and deleting redundant features. The RFI-BWO technique was subsequently utilized to extract pertinent features from the pre-processed dataset. The RFI-BWO successfully discerned characteristics pertinent to COVID-19 detection. Subsequent to feature extraction, the LLRGO algorithm was employed for feature selection, systematically selecting and assessing subsets of characteristics to determine those most pertinent to COVID-19 identification. Ultimately, the chosen features were input into the CCNN, which acquired intricate patterns and correlations within the feature set. The CCNN facilitated precise classification of COVID-19 categories through the analysis of the chosen features. The proposed technique demonstrates enhancements of 19.13% in accuracy, 9.41% in precision, and 27.18% in F1-score, and specificity relative to current methodologies. Clarifying the judgments made by the proposed model is essential for fostering trust and comprehending its rationale. Integrating explainable AI methodologies, including feature significance

References

[1] Kavitha, K. S., and Megha P. Arakeri. "Computer Vision and Machine Learning-Based Techniques for Detecting the Safety Violations of COVID-19 Scenarios: A Review." Computational Vision and Bio-Inspired Computing: Proceedings of ICCVBIC 2021 (2022): 239-251.

[2] Magazzino, Cosimo, Marco Mele, and Mario Coccia. "A machine learning algorithm to analyse the effects of vaccination on COVID-19 mortality." Epidemiology & Infection 150 (2022): e168.

[3] Absar, Nurul, et al. "The efficacy of deep learning based LSTM model in forecasting the outbreak of contagious diseases." *Infectious Disease Modelling* 7.1 (2022): 170-183.

[4] Deb, Sagar Deep, et al. "CoVSeverity-Net: an efficient deep learning model for COVID-19 severity estimation from Chest X-Ray images." Research on Biomedical Engineering 39.1 (2023): 85-98.

[5] Jain, Anurag, et al. "Iot & ai enabled three-phase secure and non-invasive covid 19 diagnosis system." Computers, Materials and Continua (2022): 423-438.

[6] Singh, Bhupinder, and Ritu Agarwal. "Coronavirus Pandemic: A Review of Different Machine Learning Approaches." Intelligent Data Communication Technologies and Internet of Things: Proceedings of ICICI 2021 (2022): 251-263.

[7] da Silva, Ana Clara Gomes, et al. "Machine learning approaches for temporal and spatio-temporal covid-19 forecasting: A brief review and a contribution." Assessing COVID-19 and Other Pandemics and Epidemics using Computational Modelling and Data Analysis (2022): 333-357.

[8] Iwendi, Celestine, et al. "COVID-19 health analysis and prediction using machine learning algorithms for Mexico and Brazil patients." Journal of Experimental & Theoretical Artificial Intelligence (2022): 1-21.

[9] Ahmed, Imran, Abdellah Chehri, and Gwanggil Jeon. "A Sustainable Deep Learning-Based Framework for Automated Segmentation of COVID-19 Infected Regions: Using U-Net with an Attention Mechanism and Boundary Loss Function." Electronics 11.15 (2022): 2296.

[10] Bayram, Fatih, and Alaa Eleyan. "COVID-19 detection on chest radiographs using feature fusion based deep learning." Signal, image and video processing 16.6 (2022): 1455-1462.

[11] Arowolo, Micheal Olaolu, et al. "Machine learning-based IoT system for COVID-19 epidemics." Computing 105.4 (2023): 831-847.

[12] Ibrahim, Zurki, Pinar Tulay, and Jazuli Abdullahi. "Multi-region machine learning-based novel ensemble approaches for predicting COVID-19 pandemic in Africa." Environmental Science and Pollution Research 30.2 (2023): 3621-3643.

[13] Zong, Kai, and Cuicui Luo. "Reinforcement learning based framework for COVID-19 resource allocation." Computers & Industrial Engineering 167 (2022): 107960.

[14] Ahmed, Imran, Gwanggil Jeon, and Abdellah Chehri. "An IoT-enabled smart health care system for screening of COVID-19 with multi layers features fusion and selection." Computing (2022): 1-18.

[15] Almagrabi, Alaa Omran, et al. "A reinforcement learning-based framework for crowdsourcing in massive health care Internet of Things." Big data 10.2 (2022): 161-170.