WJCMS

Wasit Journal of Computer and Mathematics Science

Journal Homepage: https://wjcm.uowasit.edu.iq/index.php/WJCM

e-ISSN: 2788-5879 p-ISSN: 2788-5879

Early stage prediction of COVID-19 Using machine learning model

Mohammad Abood Kadhim[1,*] and Abdulkareem Merhej Radhi[1]

[1]Computer Department/College of Science University AL-Nahrain, Baghdad, 10001, Iraq

*Corresponding Author: Mohammad Abood Kadhim

DOI: https://doi.org/10.31185/wjcm.107

Received: December 2022; Accepted: February 2023; Available online: March 2023

ABSTRACT: The healthcare sector has traditionally been an early use of technological progress and has achieved significant advantages, especially in machine learning, like predicting diseases. The COVID-19 epidemic still impacts every facet of life and necessitates a fast and accurate diagnosis. Early detection of COVID-19 is exceptionally critical to saving the lives of human beings. The need for an effective, rapid, and precise way to reduce consultants' workload in diagnosing suspected cases has emerged. This paper presents a proposed model that aims to design and implement an automated model to predict COVID-19 with high accuracy in the early stages. The dataset used in this study considers an imbalanced dataset and converted to a balanced one using Synthetic Minority Over Sampling Technique (SMOTE). Filter-based feature selection methods and many machine learning algorithms such as K-Nearest Neighbor, Support Vector Machine, Decision Tree, Logistic Regression, and Random Forest (RF) are used in this model. Since the best classification result was achieved by using the RF algorithm, and this algorithm was optimized by tuning the hyperparameters. The optimized RF enhanced the accuracy from 98.0 to 99.5.

Keywords: COVID19, SMOTE, tuning hyperparameters, Filterbased feature selection, Random Forest, Logistic Regression, and Decision tree

1. INTRODUCTION

The healthcare field is highly diverse, and the study of it takes a lot of data. Furthermore, data handling is a critical concern in this field because prompt medical care necessitates actual prediction and the dissemination of data to specialists. The COVID-19 epidemic is responsible for the significant destruction. A diverse family of viruses called COVID-19 can cause various symptoms, from the flu and the common cold to serious respiratory problems [1]. Most infected with the virus will have mild to moderate respiratory symptoms and recover without medical attention.

On the other hand, he will become seriously ill and need medical help. Serious illness is more likely to affect the elderly and those with underlying medical disorders such as cardiovascular disease, diabetes, chronic respiratory disease, or cancer. Coronavirus disease 2019 (COVID-19) may sicken anyone and cause severe illness or death at any age [2].

Many individuals nowadays are unsure whether or not they have COVID-19. For the millions of people with allergies across the world, the question becomes difficult: allergies or COVID-19 (additionally defined as the coronavirus), maybe even the flu or a cold. Fever, headache, dry cough, and sore throat are several COVID-19 complaints that patients frequently experience. When patients exhibit COVID-19 symptoms, they must visit a doctor as soon as possible. How, then, do we determine whether they have COVID-19 or not? Therefore, it's vital to understand the key distinctions. The extensive range of symptoms associated with COVID-19 varies depending on the strain. [3].

Machine learning technologies have generated a lot of attention in the research community. When compared to conventional data categorization methods, machine learning algorithms have the potential to offer excellent classification accuracy, as shown in numerous recent articles. Since it can result in sufficient protection, accurate prediction is essential.

Depending on the learning strategy employed, the forecasts' accuracy may change. Therefore, it's critical to identify tools that can accurately predict disease.

One of the most valuable methods for producing evaluations in both real-world and research a situation is machine learning classification. Persistence is also used to assess how well different machine learning techniques classify individuals as having or not having the illness. Additionally, several classification performance indicators have been used to evaluate the efficacy of these strategies [4]. A method for generating a data model that has never been classified before so that it may be used to classify new data is called classification and prediction. The classification step of machine learning can use various algorithms, including DT, Naive Bayes, KNN, SVM, RF, and others [5].

Machine learning has been used to predict the occurrence of COVID-19 in a significant number of articles that have been published as of late. Many studies use image processing to identify COVID-19. Few articles, in contrast, focused on the symptom-based method of detection.

This study presents a proposed model that intends to develop and apply a machine learning algorithm to detect Covid-19 disease in its early stages accurately. It demonstrates that reliable medical diagnoses may be made using machine learning techniques, mainly when relying on symptoms given by patients. The need to provide healthcare of the highest caliber is growing worldwide. Information systems and electronic processing systems saw global development.

For this experiment, a dataset ("COVID-19 Symptoms and Presence") was provided, and six algorithms were utilized to produce prediction models (LR, NB, SVM, KNN, DT, and RF). Also, the best accuracy is assessed using the study's best method compared to the best in earlier studies. The research investigation makes the following contributions.

1. The evaluations of the SMOTE rebalancing strategies and identifying the best precise classifier for illness risk prediction.
2. The model can provide a prompt and precise diagnosis for those who want emergency assistance.
3. When compared to past experiments, the suggested model exhibited excellent accuracy

2. LITERATURE REVIEW

According to much research on intelligent diagnostic techniques, the world's population has been adversely affected by the acute infectious disease known as Coronavirus disease 2019 (Covid-19). Many methods and ML algorithms were developed to accurately identify this dangerous illness, as shown in Table 1.

In [6], to evaluate the probability of Covid-19, the authors employed two Machine Learning Classification Techniques, Logistic Regression and Decision Tree. They are using cross-validation. The DT algorithm findings have a 98.0 percent accuracy using cross-validation, and the LR algorithm has a 97.0 percent prediction accuracy.

In [7], the artificial neural networks, support vector machine, naive Bayes, random forest, J48 decision tree, and k-nearest neighbors are several of the machine learning techniques used by the authors. Hyperparameter optimization and 10-fold cross-validation methods are also used to get better performance. The study demonstrates that support vector machines, k-nearest neighbor artificial neural networks, and random forests outperformed other methods achieving an accuracy of 98.84 percent.

In [8], the analysis and prediction of diseases benefit greatly from machine learning. Even though COVID-19 is a newer illness, machine learning has been used to predict COVID-19. This area of inquiry is constantly being expanded. In medical science, machine learning research is also being done to estimate disease development and identify viruses' existence. Using the Logistic Regression model, the technique in this work is to determine if a patient is at risk for COVID-19. The accuracy of the categorization method was 92%.

In [9], Machine learning techniques are used to create various prediction models, and their efficiency is calculated and assessed. SVM, KNN, Decision Tree Classifier, Gaussian Naive Bayesian Classifier, Multilinear Regression, Logistic Regression, XGBoost Classifier, and Random Forest are used. According to the results, the Random Forest Classifier and the DT outperform the other ML Models with an accuracy rate of 97%.

In [10], Using epidemiological labeled datasets for both positive and negative COVID-19 cases in Mexico, several machine learning methods are employed to forecast COVID-19. These techniques include artificial neural networks, decision trees, logistic regression, naive Bayes, and support vector machines. The models were trained on the training dataset, which comprised 80% of the dataset, then tested on the remainder, 20%. The decision tree algorithm was the most accurate of all models created, with an accuracy rate of 94.99 percent, when compared to other models created using logistic regression, naive Bayes, SVM, and ANN, which had accuracy rates of 94.41 percent, 94.36 percent, 92.40 percent, and 89.20 percent, respectively.

In [11], The severity of Coronavirus Disease 2019 (COVID-19) in clinical diagnosis may be distinguished and specified using the efficient, intelligent prediction model suggested in this study. The Suggested machine learning methods, like a

support vector machine model improved by a slime mold algorithm (SMA). According to the findings, the suggested SMA produces superior classification performance by 90%.

In [12] Additionally, utilizing GHS, CD3 %, total protein, and patient age as the key feature identification methods, investigators discovered four significant medical feature combinations combining clinical, labs, and demographic data. The empirical results demonstrate the effectiveness and robustness of the new model in predicting patients in critical/severe situations, with a combination of the four features producing an accuracy of 77.5%. The model's importance and supplementary resources for the medical consultant were disclosed by the survival and the cox-multivariant regression analysis.

In [13] suggested a method to identify COVID-19 and determine if it is COVID-19 or pneumonia. Rapid clinical diagnosis is difficult to achieve. As a result, a Random forest-based ML model with 95.95% accuracy is used. The dataset contains 253 records from 169 suspected patients. There are 49 characteristics per record. Only 11 variables, though, were selected as the final indications. Various validation procedures were employed to guarantee reliability.

Table 1. Literature Review

Id	Author	Year	Algorithm	Accuracy
1	A. Arista	2022	DT	0.98
2	Villavicencio, J. J. Macrohon, X. A. Inbaraj, J. H. Jeng, and J. G. Hsieh	2022	RF	0.98
3	H. Turabieh and W. Ben	2021	CNN	0.76
4	S. G. Annwesha Banerjee	2021	LR	0.92
5	K. B. Prakash, "Analysis	2020	DT, RF	0.97
6	L. J. Muhammad, E. A. Algehyne	2021	DT	0.94
7	P. Wu et al.	2021	SMA	0.90
8	This work		RF	99.5

3. METHODOLOGY

This research aims to develop a model that can predict COVID-19 with high accuracy. This section comprises a flow chart, evaluation matrices, data collection, dataset description, data pre-processing, feature engineering, and suitable machine learning algorithms. It also includes a technique and procedure section, as shown in Fig. (1), which explains our methodology. The COVID-19 training data set with 21 various characteristics is used to test the performance and validation of the developed model using several models. Confusion matrix measurements are used to estimate results, and many approaches are used to judge accuracy. To assess how well a machine learning system performed, the researcher used a train-test split. The hold-out approach for training a machine learning model uses 70-30, 70 percent for training and 30 percent for testing the machine learning algorithms.

3.1 DATASET DESCRIPTION

We used a dataset called "COVID-19 Symptoms and Presence" from Kaggle for the data collection [14]. The dataset had 21 characteristics, of which 20 were potential risk factors for contracting the virus. The final attribute, which decides whether COVID-19 is present in the sample or not, had one value out of 20. In all, 5434 rows make up the dataset. Lists the dataset's characteristics as well as its descriptions in Table (2). As demonstrated in Fig.(2), the dataset is significantly unbalanced. The collection contains 5434 samples altogether, of which 4383 are COVID-19 positive samples, and 1051 are COVID-19 negative samples.

3.2 IMBALANCED DATA

A machine learning algorithm may not learn as accurately if it is trained with unbalanced data. To create an adequate model, the imbalanced data must first be addressed using SMOTE [15] "Synthetic Minority Oversampling Technique" (SMOTE) achieves oversampling in the minority dataset by increasing the number of examples for the minority group by producing additional synthetic samples based on a predetermined number of a random sample's neighbors [16]. The dataset after converting from an imbalanced to a balanced dataset, is shown in Fig.(3).

3.3 FEATURE SELECTION

The goal of the feature selection process is to select useful variables. The quality of the data is unaffected by the removal of redundant or irrelevant characteristics since they provide contradictory information. When building a predictive model, the feature selection offers three key benefits: the chance to better model interpretation; a reduction in overfitting; and

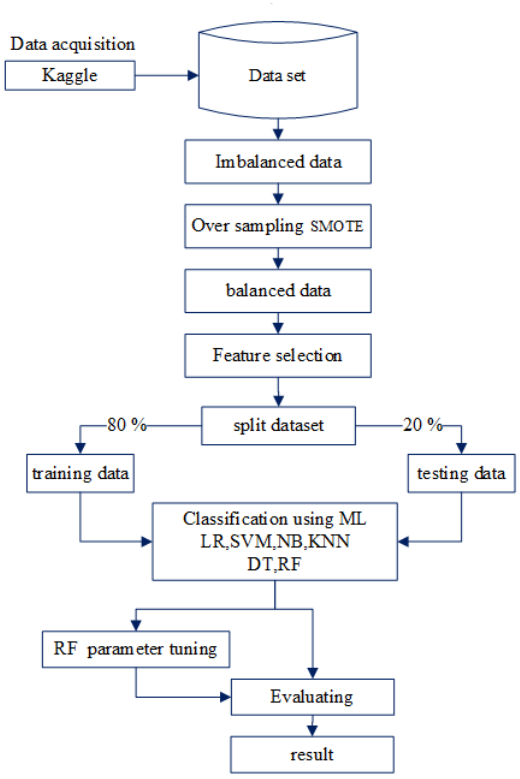

FIGURE 1. System architecture

Table 2. Attribute descriptions of the COVID-19

Attribute Name	Description
Breathing problems	suffering from breathlessness and breathing issues.
Asthma	This person has asthma.
Dry cough	persistent cough without phlegm.
Chronic lung disease	The patient has lung disease
Runny nose	The patient's nose is running.
Fever	There is an unusually heavy body temperature.
Sore throat	discomfort and a scratchy sensation in the mouth.
Heart disease	The patient has heart disease
Headache	pain in the head may be both widespread and localized
Hypertension	The patient has diabetes and high blood pressure
Fatigue	feeling worn out and feeble all the time
Diabetes	The patient has diabetes
Contact with COVID-19 patient	Contact with individuals who are COVID-19-positive
Attended large gathering	A recent large gathering was attended by the individual or anyone from the family.
Gastrointestinal	issues with the digestive system
Visited public exposed places	malls, temples, and other public locations recently visited
Family working in public	open locations Family members are employed in a busy area, a hospital, or a market.
Abroad travel	has recently traveled
Sanitation from market	Before using things you bought at the store, disinfect them.
Wearing masks	putting on face masks correctly.
COVID-19	COVID-19 is present.

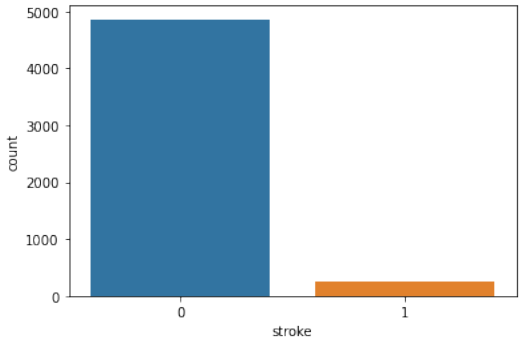

FIGURE 2. Imbalanced dataset

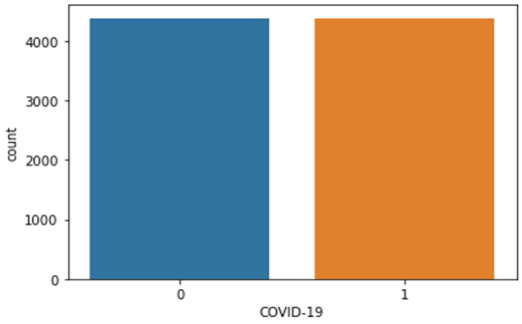

FIGURE 3. Balanced data

shorter training times, which enhance generalizations [17] [18].

3.4 FILTER-BASED FEATURE SELECTION

To determine the significance of a feature for inclusion in the subset of features, the significant properties of the data are employed. This method can be divided into two groups: Rank Based and Subset Evaluation Based. Without taking into account how different features relate to one another, rank-based categories employ simple univariate statistical approaches to rate each unique feature. [19]. choosing a feature subset that is highly correlated with classes, has little redundancy, and is suitable. Various objective functions based on information-theoretic measurements and correlation with the class attribute, are used to attain the goal. From the best feature subset, the underperforming features are eliminated [20]. In this study, the correlation method is used, 2 features Wearing Masks and Sanitization from the market, show that the correlation with the target is 0. this feature is eliminated as shown in fig 4.

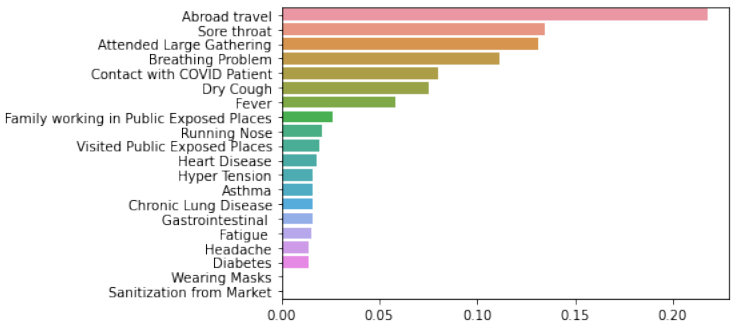

FIGURE 4. Correlation of features with the target

3.5 MACHINE LEARNING

ML is a field of artificial intelligence (AI) where data can be organized to understand the user. Computers are programmed with data, using training data or past experiences, learn parameters needed to improve computer programs, and can predict the future using this data. The primary goal of ML is to automatically learn from the input data (experience) and produce the desired output by looking for trends and patterns in the data without human intervention. [21]. machine learning techniques are divided into four kinds of learning Supervised, Unsupervised, Semi-Supervised, and Reinforcement [22]. In this work, Supervised Learning is used.

3.6 PERFORMANCE METRICS

The number of judgments that a prediction system agreed with a specialist compared to the number of disagreements with the specialist is known as a confusion matrix shown in Fig.(5). used for evaluating the performance of ML algorithms and calculating various measurements such as sensitivity, specificity, accuracy, and precision, as shown in table (3).

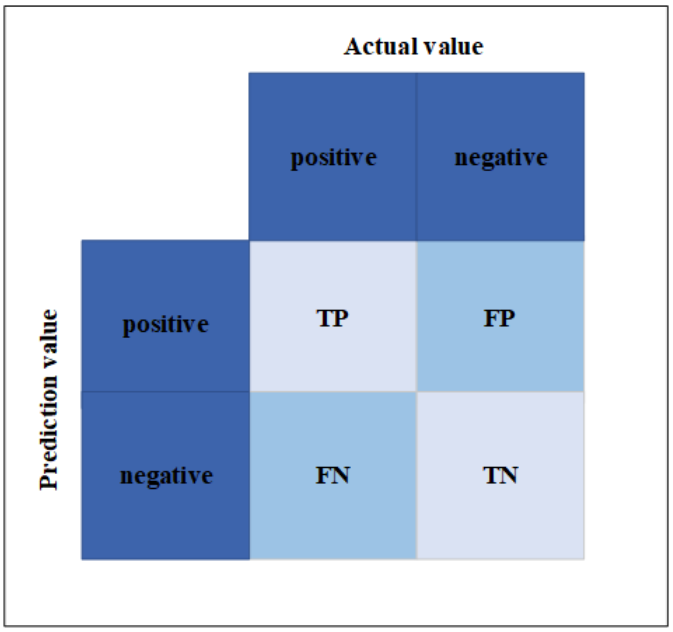

FIGURE 5. Confusion matrix

Table 3. Evaluation metric

Metric Name	Calculation
Recall	$TP/(TP + FN)$
Precision	$TP/(TP + FP)$
F1 Score	$2 * (Precision * Recall)/(Precision + Recall)$
Accuracy	$(TP + TN)/(TP + FN + TN + FP)$

3.6.1. Logistic Regression (LR):

It is a well-liked standard baseline reference approach for binary classification issues (problems with two class values) and employs a weighted combination of the input characteristics before passing them via a Sigmoid function shown in Equation. (1). The sigmoid function transforms any real value entered into a number between 0 and 1. The core of the strategy is the logistic function, sometimes known as "logistic regression." [23].

$$sigmoid = \frac{1}{1 + e^{-x}} \tag{1}$$

3.6.2. Support Vector Machines

this method is used to analyze data and discover patterns in classification and regression analysis [24]. SVM is built on identifying an optimal hyperplane that splits a data set into two subclasses in the best possible manner. The distance between the hyperplane and the data set's nearest point is called The margin.

3.6.3. K-Nearest Neighbor

It is a categorization system based on distance measurements. It is an instance-based categorization, which implies that comparable instances are classified similarly. The slow or lazy algorithm is another name for it. We have the appropriate X-value and Y-value for each point. Researchers receive the Y-value of both instances when we are given a new instance in terms of the X-value and find the Y-value of the instance. We want to accurately estimate the majority class based on Yi values [25].

3.6.4. Naïve Bayesian based

One of the most popular ML algorithms and according to the Bayes Theorem, the likelihood of an occurrence may be described in terms of related events [26]. It is simple to construct naive Bayes classifiers without complex iterative parameters, particularly when used for medical data. Equation (2) represents this theorem as follows:

$$p\left(\frac{x}{y}\right) = \frac{p\,(y) * p\left(\frac{x}{y}\right)}{p\,(x)} \tag{2}$$

Where P(y) is the probability of y, P(x) is the probability of x, P(y/x) is the probability of y given x, and P(x/y) is the probability of x given y. the final probability is a result of chaining input features of a class [27].

3.6.5. Decision Tree

the algorithm used to predict or classify discrete or continuous data. The divide and conquer strategy was used. It lacks subject expertise and parameter setting. It addresses the massive data dimensionality. Exploring knowledge discovery is suitable. It is simpler to comprehend and understand the decision tree's outcomes [28], [29]

3.6.6. Random Forest algorithm

this algorithm is used for regression and classification to train numerous Decision Trees, create models, and then use the voting results of multiple Decision Trees. The only parameters required by the Random Forest method are the number of DT to be built (T) and the number of input characteristics to be taken into account when each Decision Tree node is divided (M) [30].

3.7 HYPERPARAMETER OPTIMIZATION

hyperparameters in ML are used to control the algorithm performance. Hyperparameter optimization checks a set of hyperparameters of the learning algorithm to enhance the performance of the ML algorithm [31]. There are different types of hyperparameters, and all algorithm has a different tuning process for these hyperparameters [32].

3.7.1. Hyperparameters in RF

RF classifiers have many hyperparameters that must be determined and applied to optimize. Table (4) shows the RF classifier's hyperparameters and default values.

Table 4. RF hyperparameters

classifier	Parameters	Default values
RF	Min- samples split	1
	n-estimator	10
	Max depth	None
	Bootstrap	Ture

4. CLASSIFICATION STAGE AND RESULT

Using the COVID-19 Symptom's training dataset with 21 distinct features, several ML algorithms are utilized to assess the performance and validity of the created model. Confusion matrix measurements were used to estimate results, and

different techniques were used to evaluate the accuracy. A train-test split validation technique was used to split the dataset into a ratio of 80% for training and 20% for testing and assessment. The models were constructed using the default parameters, and the performance was evaluated based on the unseen test. Table (5) shows the classification results of different algorithms applied to the dataset which RF has the highest classification result compared to other algorithms, followed by DT. Fig. (6) shows the confusion matrix of all classification algorithms used.

Table 5. Classification result

Algorithm	Acc	Precision	Recall	specificity	f1_score
LR	95.0	0.97	0.94	0.97	0.95
SVM	98.0	1.0	0.97	1.0	0.98
KNN	98.0	1.0	0.96	1.0	0.98
NB	83.0	1.0	0.66	1.0	0.80
DT	98.0	1.0	0.96	1.0	0.98
RF	98.0	1.0	0.96	1.0	0.98

4.1 OPTIMIZATION STAGE AND RESULT:

Since the RF algorithm was the highest results, a second part of the experiment was done, which is the process of tuning the parameters of the RF algorithm, which is a randomized search used to identify the best parameters, as shown in table (5) from many parameters in RF. Table (6) shows the values of RF hyperparameters after tuning, which affect the results of classification as shown in table (7), while table (8) shows the comparison between the previous and this work.

Table 6. RF hyperparameters after tuning

Classifier	Parameter	value
RF	Min- samples split	10
	n-estimator	20
	Max depth	1000
	Bootstrap	False

Table 7. Optimization result

Algorithm	Acc	Precision	recall	specificity	f1_score
RF	99.5	1.0	0.96	1.0	0.98

Table 8. Comparison with pervasive works

Studies	Classifier	Best Acc
[6]	DT	98.0
[7]	RF	98.0
This	RF	98.95%
work	RF+ Tuning parameters	99.5

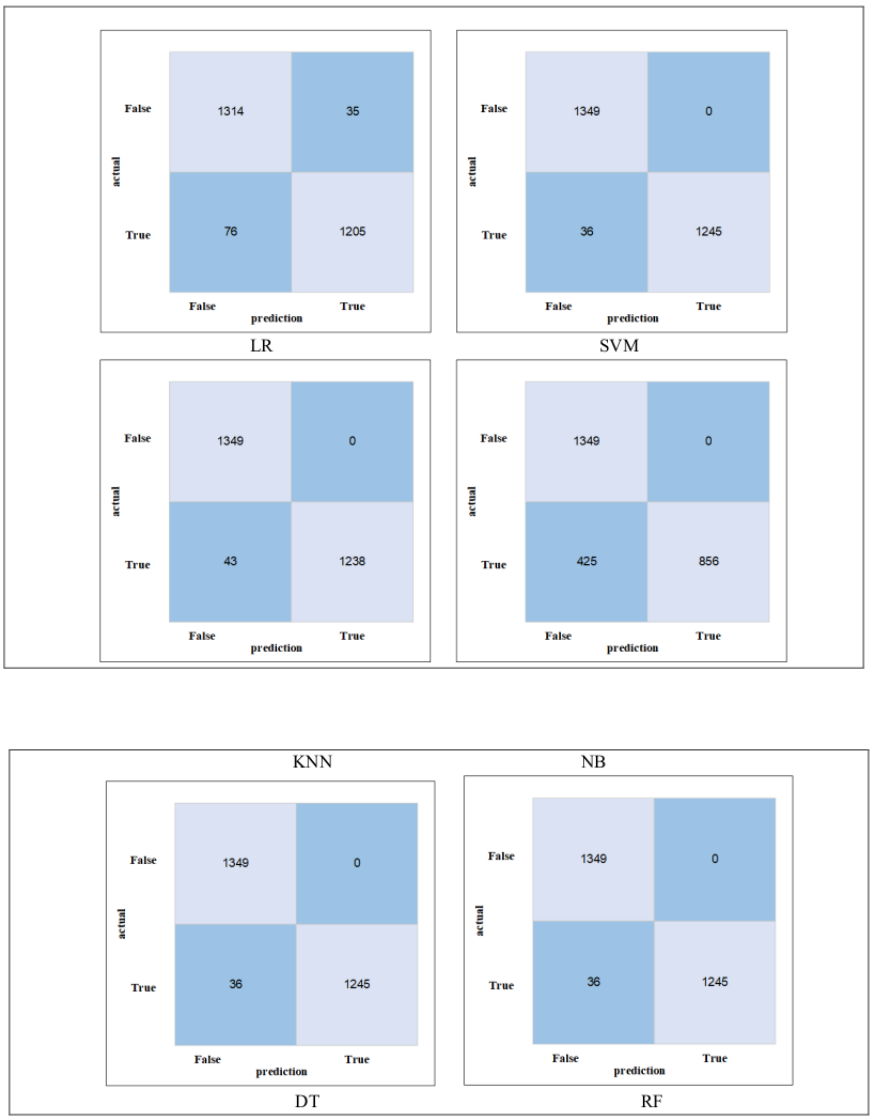

FIGURE 6. Confusion matrix of all classification algorithms used

5. CONCLUSIONS

In this work, the imbalanced dataset problem was solved using SMOTE methods, producing several features. Filter-based feature selection was used to select the important features when using this method to conclude that (Wearing Masks and Sanitization from the market because the correlation with the target is (0) were unimportant features. Using ML algorithms for classification, RF shows the best classification result among other algorithms. In the RF c algorithms, many hyperparameters must choose the best for tuning to enhance the classification result, such as Min- samples split, n-estimator, Max depth, and Bootstrap, which helps to optimize the accuracy result from 98.0 to 99.5.

FUNDING

None

ACKNOWLEDGEMENT

None

CONFLICTS OF INTEREST

The author declares no conflict of interest.

REFERENCES

[1] E. Gambhir, R. Jain, A. Gupta, and U. Tomer, "Regression analysis of COVID-19 using machine learning algorithms," *2020 International conference on smart electronics and communication (ICOSEC)*, pp. 65–71, 2020.

[2] "World Health Organization . Coronavirus 2021," 2022. https://www.who.int/health-topics/coronavirus(accessed2022.

[3] L. Wynants, "Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal," *BMJ*, vol. 369, pp. 1328–1328, 2020.

[4] A. K. Dwivedi, "Performance evaluation of different machine learning techniques for prediction of heart disease," *Neural Computing and Applications*, vol. 29, no. 10, pp. 685–693, 2016.

[5] S. S. P. Shimpi, M. Shroff, and A. Godbole, "A Machine Learning Approach for the lassification of Cardiac Arrhythmia," *2017 International Conference on Computing Methodologies and Communication (ICCMC)*, 2017.

[6] A. Arista, "Comparison Decision Tree and Logistic Regression Machine Learning Classification Algorithms to determine Covid-19," *Sinkron*, vol. 7, no. 1, pp. 59–65, 2022.

[7] C. N. Villavicencio, J. J. Macrohon, X. A. Inbaraj, J. H. Jeng, and J. G. Hsieh, "Development of a Machine Learning Based Web Application for Early Diagnosis of COVID-19 Based on Symptoms," *Diagnostics (Basel)*, vol. 12, no. 4, 2022.

[8] S. G. A. B. Majumder and D. Singh, "An Intelligent System for Prediction of COVID-19 Case using Machine Learning Framework-Logistic Regression," *Journal of Physics*, pp. 2021–2021.

[9] K. B. Prakash, "Analysis, Prediction and Evaluation of COVID-19 Datasets using Machine Learning Algorithms," *International Journal of Emerging Trends in Engineering Research*, vol. 8, no. 5, pp. 2199–2204, 2020.

[10] L. J. Muhammad, E. A. Algehyne, S. S. Usman, A. Ahmad, C. Chakraborty, and I. A. Mohammed, "Supervised Machine Learning Models for Prediction of COVID-19 Infection using Epidemiology Dataset," *SN Comput Sci*, vol. 2, no. 1, pp. 2021–2021.

[11] P. Wu, "An Effective Machine Learning Approach for Identifying Non-Severe and Severe Coronavirus Disease 2019 Patients in a Rural Chinese Population: The Wenzhou Retrospective Study," *IEEE Access*, vol. 9, pp. 45486–45503, 2021.

[12] J. Cao, Z. Zhang, J. Du, L. Zhang, Y. Song, and G. Sun, "Multi-geohazards susceptibility mapping based on machine learning-A case study in Jiuzhaigou, China," *Natural Hazards*, vol. 102, no. 3, pp. 851–871, 2020.

[13] J. Wu, "Rapid and accurate identification of COVID-19 infection through machine learning based on clinical available blood test results," 2020.

[14] C. Symptoms and Presence. https://www.kaggle.com/datasets/hemanthhari/symptoms-and-covid-presence.

[15] D. Dablain, B. Krawczyk, and N. V. Chawla, "DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data," *IEEE Trans Neural Netw Learn Syst*, 2022.

[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[17] K. Anitha, "Rough neural network," *Asian Journal of Research in Social Sciences and Humanities*, vol. 6, no. cs1, pp. 413–421, 2016.

[18] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, 2014.

[19] U. M. Khaire and R. Dhanalakshmi, "Stability of feature selection algorithm: A review," *Journal of King Saud University-Computer and Information Sciences*, 2019.

[20] K. Jha and S. Saha, "Incorporation of multimodal multiobjective optimization in designing a filter based feature selection technique," *Applied Soft Computing*, vol. 98, pp. 106823–106823, 2021.

[21] C. Rao and V. N. Gudivada, "Computational analysis and understanding of natural languages: principles, methods and applications," 2018. Elsevier.

[22] N. Abuja, "Prediction Of Heart Disease Using Bayesian Network Model," 2019.

[23] D. Namly, K. Bouzoubaa, A. E. Jihad, and S. L. Aouragh, "Improving Arabic lemmatization through a lemmas database and a machine-learning technique," *Recent Advances in NLP: The Case of Arabic Language*, pp. 81–100, 2020.

[24] V. Sharma, S. Yadav, and M. Gupta, "Heart Disease Prediction using Machine Learning Techniques," *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, 2020.

[25] P. A. T. Azhar and M, "Comparative Review of Feature Selection and Classification modeling," *presented at the 2019 International Conference on Advances in Computing, Communication and Control (ICAC3)*, vol. 1, 2019.

[26] J. Yu, S. Greco, P. Lingras, G. Wang, and A. Skowron, "Rough set and knowledge technology: 5th international conference, rskt 2010," Springer, 2010.

[27] K. Vembandasamy, R. Sasipriya, and E. Deepa, "Heart diseases detection using Naive Bayes algorithm," *International Journal of Innovative Science, Engineering & Technology*, vol. 2, no. 9, pp. 441–444, 2015.

[28] S. Vijiyarani and S. Sudha, "Disease prediction in data mining technique-a survey," *International Journal of Computer Applications & Information Technology*, vol. 2, no. 1, pp. 17–21, 2013.

[29] J. Soni, U. Ansari, D. Sharma, and S. Soni, "Predictive data mining for medical diagnosis: An overview of heart disease prediction," *International Journal of Computer Applications*, vol. 17, no. 8, pp. 43–48, 2011.

[30] G. Biau and E. Scornet, "A random forest guided tour," *test*, vol. 25, no. 2, pp. 197–227, 2016.

[31] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice," *Neurocomputing*, vol. 415, pp. 295–316, 2020.

[32] M. Feurer and F. Hutter, "Hyperparameter optimization," in *Automated machine learning*, pp. 3–33, Springer, 2019.