International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 10 Issue: 09 | Sep 2023 www.irjet.net p-ISSN: 2395-0072

Symptom-Based Prediction of COVID-19 using Machine Learning Models with SMOTE for Class Imbalance

Hari Priya. N[1], Dr.S. Rajeswari[2]

[1]Research Scholar, Department of Computer Science, Sree Saraswathi Thyagaraja College, Tamil Nadu, India
[2]Associate Professor, Department of Computer Science, Sree Saraswathi Thyagaraja College, Tamil Nadu, India

---

Abstract - *The global dissemination of Coronavirus Disease-19 (COVID-19) has emerged as a significant peril, threatening the lives of numerous individuals. The timely detection and precise identification of those who have been affected are of paramount importance in the realm of disease control. The study utilizes the COVID-19 Open Research Dataset Challenge (CORD-19) dataset, which was obtained from Kaggle. This dataset consists of 127 cases and 20 features. The proposed approach encompasses a data pre-processing stage, followed by the use of a feature selection strategy utilizing the wrapper method to find the most significant features. We used the SMOTE (Synthetic Minority Over-sampling Technique) technique to address the class imbalance issue, as the dataset used in this study contains unbalanced values in the target variable. The output from the SMOTE method was fed into five different machine learning algorithms such as Random Forest, SVM, Logistic Regression, Naive Bayes, and XGBOOST for classification. Then the machine learning algorithms were tuned by parameter optimization method to achieve higher accuracy. Based on experimental data, the Random Forest classifier combined with SMOTE outperformed all other classifiers with an accuracy of 98%.*

***Key Words*: COVID-19, Machine Learning, Classification, Feature Selection, Prediction, SMOTE**

1.INTRODUCTION

COVID-19 (Coronavirus disease 2019) is a highly infectious respiratory illness caused by the novel coronavirus SARS-CoV-2. The novel coronavirus-based infectious disease was given the official name COVID-19 (coronavirus disease 2019) by WHO in February 2020, and later it was declared as pandemic in March 2020 [1]. COVID-19 is primarily transmitted through respiratory droplets when an infected person speaks, coughs, or sneezes, but it can also be transmitted by touching a contaminated surface and then touching one's face. The symptoms include fever, cough, shortness of breath, fatigue, body aches, loss of taste or smell, sore throat, congestion, and diarrhea, and can range from mild to severe.

The global impact of the COVID-19 pandemic has been unprecedented, affecting millions of people worldwide. The pandemic has wreaked havoc on the economy and society, resulting in mass layoffs, reduced productivity, and increased poverty. Public health measures such as social isolation, using masks, and frequent hand washing are critical in reducing the virus's spread. In addition to these public health measures, COVID-19 vaccines have been developed and are being distributed globally. These vaccines were developed and tested in record time, and they were found to be highly effective in preventing COVID-19 infection and reducing disease severity [2].

COVID-19 prediction using Machine Learning (ML) entails analysing pandemic data and predicting future outcomes using advanced algorithms and models. Machine learning models can identify patterns and predict future trends by analysing large datasets of COVID-19 cases, hospitalization rates, mortality rates, symptom data, and clinical data. To analyse the symptoms data and develop predictive models, various ML techniques can be used. These models can be trained using symptom data and then used to predict the likelihood of COVID-19 infection [3].

The proposed system includes pre-processing the data to handling missing values, eliminating redundant values and selecting the most informative features. After pre-processing, we used feature selection technique to identify the most crucial features using wrapper method known as recursive feature elimination (RFE). As the dataset used in this study has significantly more instances of one class (majority class) than the other class (minority class), SMOTE technique is used to handle the class imbalance problem. With the results of the SMOTE technique, five different machine learning such as Random Forest, SVM, Logistic Regression, Naïve Bayes and XGBOOST was used for classification. The performance of the machine learning models can be enhanced by the adoption of parameter optimization technique. Grid search CV has been used for optimizing the performance of machine learning models considered for this study. Also, the results of these different models were compared with and without SMOTE technique.

2. BACKGROUND STUDY

Using a machine learning (ML) data-driven method, Izquierdo et al. [4] devised a model to predict ICU admission. In order to conduct the study, a data collection of 10,504 COVID-19 patients from the general Castilla-La Mancha (Spain) population were employed, of whom 1353 were hospitalized and 83 were admitted to the intensive care unit. The data set contained clinical details about the diagnosis, severity, and outcome of the infection. It was done using a Decision Tree algorithm. Accuracy, recall, and AUC values for the model were 0.68, 0.71, and 0.76, respectively. Age, fever, and tachypnea with or without respiratory crackles were the three factors that most strongly predicted ICU admission.

Nemati et al. [5] developed a prediction model using the clinical data of the patients to calculate the length of hospital stay for COVID-19 patients. An open-access data set compiled by a team of researchers from several institutions and research labs had a data set of 1182 hospitalized patients. Several survival analysis models were implemented using a variety of statistical analysis techniques and ML strategies. With a C-index of 71.47, the stagewise gradient-boosting survival model produced the best accurate discharge-time estimate. The results showed that males and older age groups had reduced discharge probabilities.

On the basis of nine easy survey questions, Shoer et al. [6] created a prediction model. The study made use of a set of data from an Israeli national symptom survey that received over two million responses. There were 43,752 adults in all, and 498 of them self-reported having COVID-19 positive status. Age, gender, previous medical issues, smoking habits, and self-reported symptoms such as fever, sore throat, cough, shortness of breath, and loss of taste or smell were all included in the study. The model had an AUC score of 0.737 after being trained with Logistic regression approach.

Yazeed et al. [7] developed a machine learning method that was trained using data from 51,831 tested people, of whom 4769 had COVID-19, according to the records. Only eight binary features, including as sex, age, known contact with an infected person, and the emergence of five early clinical symptoms, were used by the algorithm to predict COVID-19 test results with high accuracy. The authors developed a model that diagnoses COVID-19 instances based on fundamental parameters retrieved by asking simple questions, using data collected across the country that was publicly disclosed by the Israeli Ministry of Health. The authors concluded that the methodology can be used to prioritize COVID-19 testing when testing resources are scarce.

From the existing studies that were available for predicting COVID -19 using ML techniques based on symptoms [6,7], they haven't used feature selection technique to select the best features and the class imbalance problem has not been considered. In this study, feature selection has been adopted to find the best feature set and to identify the most alarming symptoms of COVID-19 positive cases and class imbalance problem has been solved by using SMOTE technique.

3. METHODOLOGY

In this section, the overall process flow of the proposed method is shown in Figure 1. The process flow starts with Covid-19 data collection followed by pre-processing of the data. Then the processed data is applied with feature selection technique before training the model. The obtained feature subsets are used by SMOTE technique and then the outcome is used by machine learning algorithm which was optimized with hyperparameter technique and finally evaluation metrics of these classifiers were calculated.

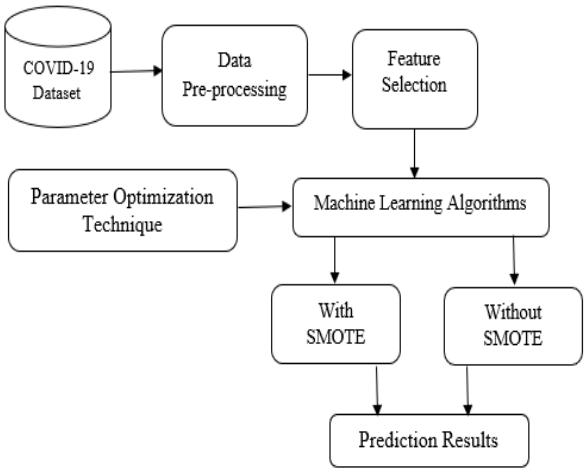

Fig - 1: Proposed Workflow

3.1. DATASET DESCRIPTION

The dataset used in this study belongs to COVID-19 Open Research Dataset Challenge (CORD-19) which has been downloaded from Kaggle [8]. It comprises of 127 instances of patient's data with 20 attributes. The features of the dataset are Age, Gender, Body Temperature, Dry Cough, Sore Throat, Weakness, Breathing Problem, Drowsiness, Fever, Travel History to Infected Countries, Diabetes, Heart Disease, Lung Disease, Stroke or Low Immunity, High BP, Kidney Disease, Change in Appetite, Loss of Smell and Target (COVID -19 positive or negative). At the pre-processing stage, we removed redundant data elements and missing values are handled by mean imputation method.

Histograms: These are used to provide insights into the distribution of age and body temperature:

Age Distribution:

The majority of the individuals in the dataset are between 20 and 40 years old. There's a smaller number of older individuals, especially above 60 which is shown in Figure 2.

Body Temperature Distribution:

Most individuals have body temperatures close to the average human body temperature (around 98.6°F).

There's a notable number of individuals with elevated temperatures, reaching up to around 103°F, which can be indicative of fever which is depicted in Figure 3.

Correlation Heatmap:

Variables like breathing problem, drowsiness, and pain in chest have a relatively higher positive correlation with the target. This suggests that these symptoms might be strong indicators or predictors for the target variable.

The travel history to infected countries variable also shows a notable correlation with the target, indicating that travel history might play a significant role in determining the outcome. Most of the symptoms show some level of positive correlation with each other, suggesting that the presence of one symptom might increase the likelihood of another and the complete correlation heatmap for the entire dataset is shown in Figure 4.

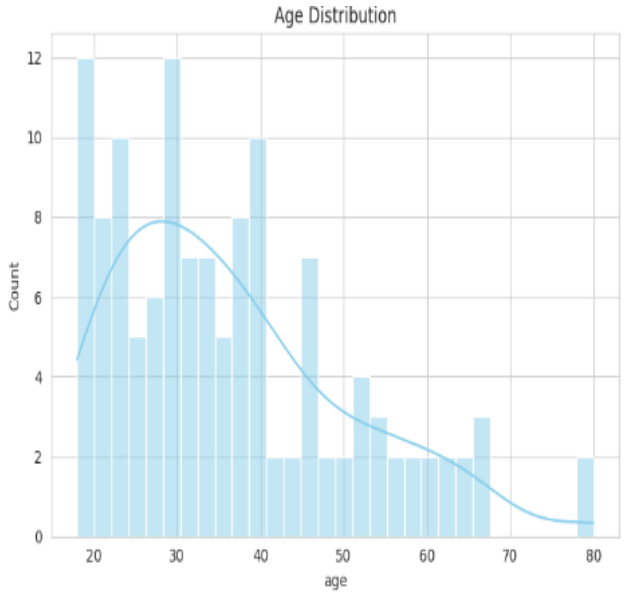

Fig - 2: Age Distribution

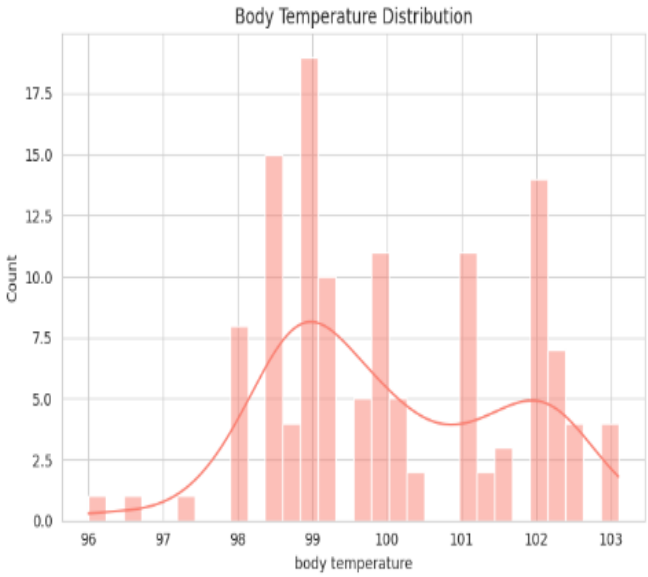

Fig - 3: Blood Temperature Distribution

3.2 FEATURE SELECTION

Feature selection is critical in machine learning because it can improve model accuracy and performance, reduce overfitting and complexity of the model, and handle issues with high dimensionality [9]. The ultimate motive of using feature selection technique is to identify the most crucial features that are highly correlated with the target variable. By selecting the most relevant features, we can improve the accuracy and efficiency of machine learning models, and gain insights into the underlying patterns and relationships in the data [10].

In this study, wrapper method known as Recursive Feature Elimination (RFE) based on SVM has been used to measure the linear correlation between two variables and to identify the top most features in the dataset. It selects features based on the performance of a specific machine learning algorithm. In SVM-RFE, the SVM algorithm is used to rank and select the most important features, and then the performance of the SVM model is evaluated on the selected features. The selected subset of features is then used as an input for machine learning models.

Using Python, we performed SVM-RFE feature selection by using an estimator of Support Vector Classification (SVC) with a linear kernel. First, an instance of SVC with a linear kernel is created as the estimator, and an instance of RFE is created by specifying the number of features to select. The RFE instance is also fitted to the training data with a specified number of features. Finally, the selected features and their rankings are fetched and are stored in a Pandas Data Frame, which is sorted by rank in ascending order. Out of all the features, the most significant features that are highly related to the presence of COVID-19 has the

rank as 1 and they are shown in Figure 5. The dataset has been split into training and test sets before applying the feature selection technique.

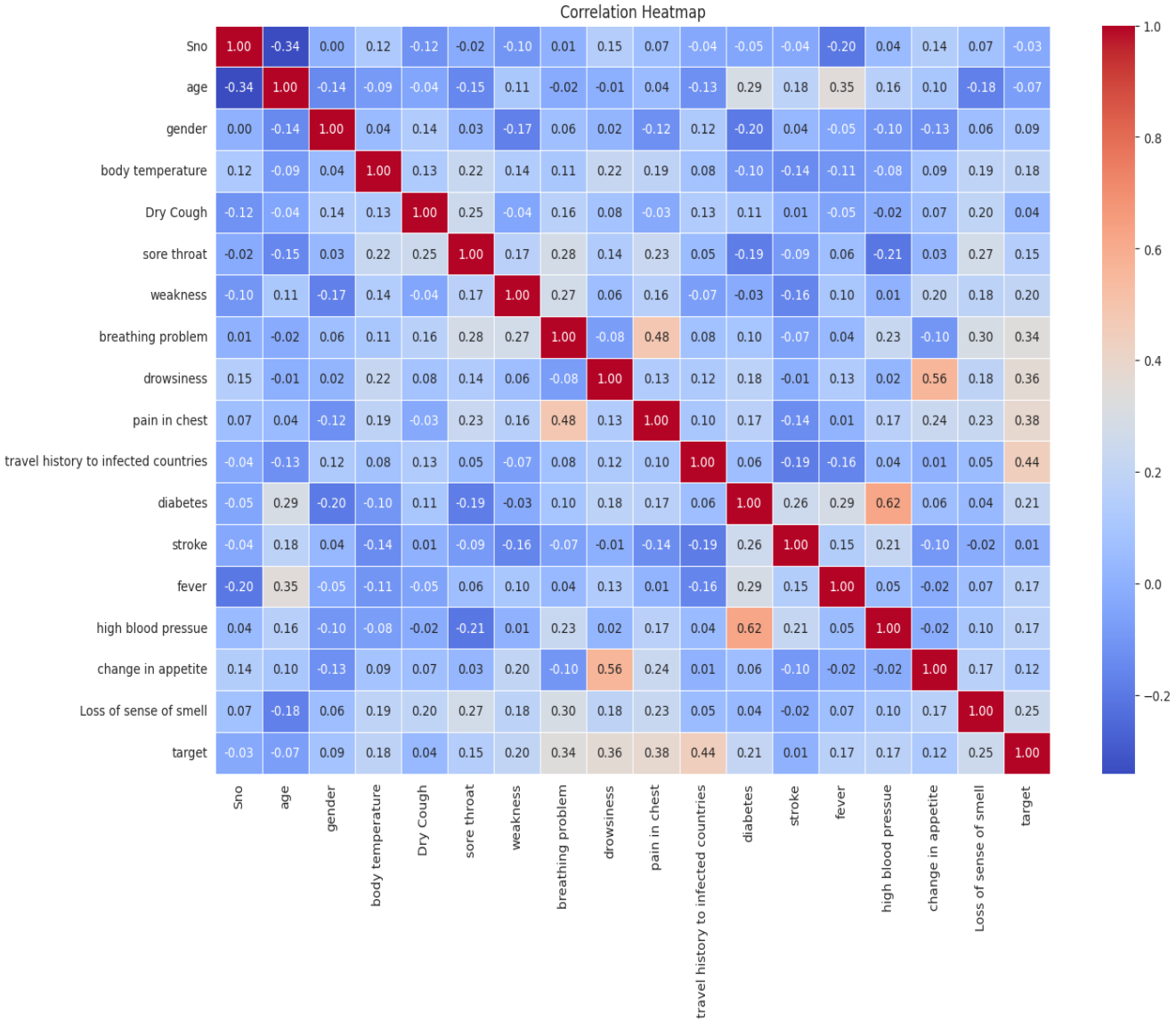

Fig - 4: Correlation Heatmap

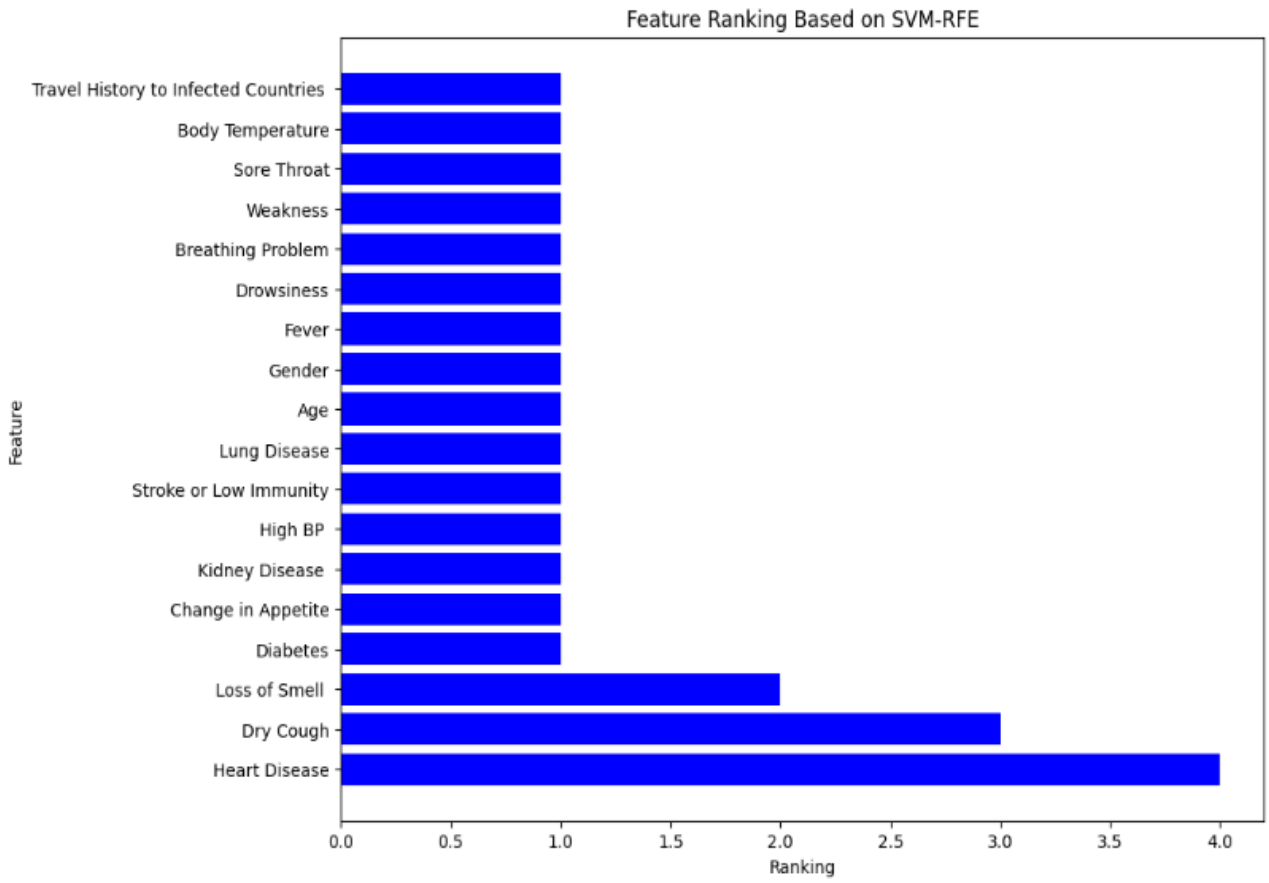

Fig - 5: Visualization of Feature Ranking based on SVM-RFE

3.3MACHINE LEARNING ALGORITHMS

SMOTE (Synthetic Minority Over-sampling Technique)

SMOTE is used in this work before training the model to address the class imbalance problem that exists in the dataset. The target variable in the dataset has 94 COVID-19 positive cases and 33 negative cases which is highly imbalanced, this problem is solved by adopting SMOTE technique.

After applying SMOTE, we built a predictive model using five machine learning algorithms after selecting the best feature subset. Random Forest, SVM, Logistic Regression, Nave Bayes, and XGBOOST are the algorithms used in this study.

Parameter Optimization Technique:

Scikit-learn offers the Grid Search CV function to do a comprehensive search across a given parameter grid. In the field of machine learning, it has been used various purposes and the most important will be these two:

Hyperparameter Optimization: Hyperparameters in machine learning models are generally predetermined and not discovered through training. The model's efficiency may be drastically altered by these values**.**

Grid Search CV: It is an abbreviation for cross-validation, which is what the "CV" stands for. Grid Search CV uses k-fold cross-validation, in which the model is trained k times on k-1 different subsets of the data and validated on the remaining subset, for each possible combination of hyperparameters. This yields a performance estimate for the model that is independent of the specific data partition used during training.

Evaluation Metrics

To compute the performance of the different ML models, we evaluated true positives (TP), False Positives (FP), True Negatives (TN), False Negatives (FN).

TP: A true positive is that a person who has Covid-19 is correctly identified as positive by a diagnostic test.

FP: A false positive would occur if a person who does not have Covid-19 is incorrectly identified as positive by a diagnostic test.

TN: A true negative would occur if a person who does not have Covid-19 is correctly identified as negative by a diagnostic test.

FN: A false negative would occur if a person who actually has Covid-19 is incorrectly identified as negative by a diagnostic test.

The different performance metrics considered in this study are listed below;

Accuracy: This is a measure of how often a classifier is correct. It is calculated as the ratio of the number of correct predictions to the total number of predictions made.

The formula for accuracy is:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN)$$

Precision: This is a measure of how many of the positive predictions made by a classifier are correct. It is calculated as the ratio of the number of true positive predictions to the total number of positive predictions made.

The formula for precision is:

$$Precision = TP / (TP + FP)$$

Recall: This is a measure of how many of the actual positive cases a classifier is able to correctly identify. It is calculated as the ratio of the number of true positive predictions to the total number of actual positive cases.

The formula for recall is:

$$Recall = TP / (TP + FN)$$

F1 Score: This is a measure of the overall accuracy of a classifier that takes both precision and recall into account. It is calculated as the harmonic mean of precision and recall.

The formula for F1 score is:

$$F1\ Score = 2 * (Precision * Recall)/ (Precision + Recall)$$

4. RESULTS AND DISCUSSION

The prediction of COVID-19 using machine learning approach is carried out based on the symptoms data. In this work, five distinct machine learning models based on classification were used with parameter optimization technique known as Grid Search CV, along with the feature selection technique, and finally SMOTE method is used before training the model. Various parameters in the dataset are evaluated and the most crucial features pertaining to the target variable are identified and the best machine learning model for COVID-19 prediction has been identified. We used Python libraries such as pandas, numpy, scikit-learn, matplotlib, and seaborn to implement this work.

Based on the results from the feature selection technique, wrapper method that uses SVM-RFE has identified the top and most important features which are highly related with the target variable. The list of identified symptoms were **t**ravel History to Infected Countries, Fever, Breathing Problem, Drowsiness, Loss of Smell, Diabetes, Weakness, Lung Disease, Body Temperature, Stroke or Low Immunity. After feature selection, we evaluated the ML algorithms using different metrics. Out of the five ML algorithms, random forest algorithm achieved highest accuracy of 98% comparing with all other algorithms. The accuracy of the same algorithm differs with the application of SMOTE technique**.** The evaluation metrics of different ML algorithms with and without SMOTE has been depicted in Table 1 and 2. From the table it is clear that SMOTE not only deals with class imbalance problem but also aids in improving the accuracy of the classifiers**.**

Table 1: Performance Metrics of ML Algorithms (With SMOTE)

Algorithm/Measures	Accuracy
Random Forest	98%
SVM	94%
Logistic Regression	93%
Naïve Bayes	92%
XGBoost	97%

Table 2: Performance Metrics of ML Algorithms (Without SMOTE)

Algorithm/Measures	Accuracy
Random Forest	93%
SVM	91%
Logistic Regression	90%
Naïve Bayes	89%
XGBoost	92%

5. CONCLUSION

In pandemic situations, it is crucial to pinpoint those who are susceptible to infection and the spread of disease in order to develop treatment and prevention plans. In this study, the most important symptoms were identified and

class imbalance problem has been sorted by using SMOTE. According to experimental findings, random forest classifier performs better than other classifiers utilized in this work for COVID-19 diagnosis. When the system is overloaded due to overcrowding, this approach can assist hospitals and medical facilities in deciding which patient require immediate attention before other patients and also help avoid delays in providing the required care.

6. FUTURE ENHANCEMENTS

For the future work, incorporating additional data sources could be done as the current model uses only symptom data to predict COVID-19, but other data sources such as their medical history, location data, blood test reports could be added to improve accuracy. Implementing the model in clinical settings is another suggestion, if the model is validated and found to be accurate, it could be integrated into clinical settings to assist with COVID-19 diagnosis. As new variants of COVID-19 emerge, the model will need to be adapted to accurately predict these variants. This could involve collecting new data and training the model on the new variants, or adapting the existing model to incorporate information about the genetic mutations present in the new variants.

REFERENCES

[1] Pourhomayoun, M., & Shakibi, M. (2021). Predicting mortality risk in patients with COVID-19 using machine learning to help medical decision-making. Smart Health, 20, 100178.

[2] Asselah, T., Durantel, D., Pasmant, E., Lau, G., & Schinazi, R. F. (2021). COVID-19: Discovery, diagnostics and drug development. Journal of hepatology, 74(1), 168-184.

[3] Chang, D., Chang, D., & Pourhomayoun, M. (2019, December). Risk prediction of critical vital signs for icu patients using recurrent neural network. In 2019 International Conference on Computational Science and Computational Intelligence (CSCI) (pp. 1003-1006). IEEE.

[4] Izquierdo, J. L., Ancochea, J., Savana COVID-19 Research Group, & Soriano, J. B. (2020). Clinical characteristics and prognostic factors for intensive care unit admission of patients with COVID-19: retrospective study using machine learning and natural language processing. Journal of medical Internet research, 22(10), e21801.

[5] Nemati, M., Ansary, J., & Nemati, N. (2020). Machine-learning approaches in COVID-19 survival analysis and discharge-time likelihood prediction using clinical data. Patterns, 1(5), 100074.

[6] Shoer, S., Karady, T., Keshet, A., Shilo, S., Rossman, H., Gavrieli, A., ... & Segal, E. (2021). A prediction model to prioritize individuals for a SARS-CoV-2 test built from national symptom surveys. Med, 2(2), 196-208.

[7] Zoabi, Y., Deri-Rozov, S., & Shomron, N. (2021). Machine learning-based prediction of COVID-19 diagnosis based on symptoms. npj digital medicine, 4(1), 3.

[8] https://www.kaggle.com/datasets/bitsofishan/covid19-patient-symptoms

[9] de Moraes Batista, A. F., Miraglia, J. L., Rizzi Donato, T. H., & Porto Chiavegatto Filho, A. D. (2020). COVID-19 diagnosis prediction in emergency care patients: a machine learning approach. MedRxiv, 2020-04.

[10] Alakus, T. B., & Turkoglu, I. (2020). Comparison of deep learning approaches to predict COVID-19 infection. Chaos, Solitons & Fractals, 140, 110120.