Bulletin of Electrical Engineering and Informatics
Vol. 13, No. 3, June 2024, pp. 1718~1729
ISSN: 2302-9285, DOI: 10.11591/eei.v13i3.6774

An efficient synthetic minority oversampling technique-based ensemble learning model to detect COVID-19 severity

Smriti Mishra[1], Ranjan Kumar[2], Sanjay K. Tiwari[3], Priya Ranjan[4]
[1]Department of Computer Science, Gaya College, Gaya, India
[2]Department of Computer Science, Aryabhatta College, University of Delhi, Delhi, India
[3]Department of Mathematics, Magadh University, Bodhgaya, India
[4]School of Computer Science, UPES, Uttarakhand, India

Article Info

Article history:

Received May 22, 2023
Revised Sep 13, 2023
Accepted Sep 28, 2023

Keywords:

COVID-19 severity
Ensemble learning
Feature engineering
Machine learning techniques
Synthetic minority oversampling technique

ABSTRACT

The COVID-19 pandemic has highlighted the importance of accurately predicting disease severity to ensure timely intervention and effective allocation of healthcare resources, which can ultimately improve patient outcomes. This study aims to develop an efficient machine learning (ML) model based on patient demographic and clinical data. It utilizes advanced feature engineering techniques to reduce the dimensionality of dataset and address the issue of highly imbalanced data using synthetic minority oversampling technique (SMOTE). The study employs several ensemble learning models, including XGBoost, Random Forest, AdaBoost, voting ensemble, enhanced-weighted voting ensemble, and stack-based ensembles with support vector machine (SVM) and Gaussian Naïve Bayes as meta-learners, to develop the proposed model. The results indicate that the proposed model outperformed the top-performing models reported in previous studies. It achieved an accuracy of 0.978, sensitivity of 1.0, precision of 0.875, F1-score of 0.934, and receiver operating characteristic area under the curve (ROC-AUC) of 0.965. The study identified several features that significantly correlated with COVID-19 severity, which included respiratory rate (breaths per minute), c-reactive proteins, age, and total leukocyte count (TLC) count. The proposed approach presents a promising method for accurate COVID-19 severity prediction, which may prove valuable in assisting healthcare providers in making informed decisions about patient care.

Corresponding Author:

Ranjan Kumar
Department of Computer Science, Aryabhatta College, University of Delhi
5, Benito Juarez Road, New Delhi, Delhi-110021, India
Email: ranjan301@gmail.com

1. INTRODUCTION

In recent years, the global response to the ongoing novel coronavirus disease (COVID-19) pandemic, caused by COVID-19, has been marked by its rapid spread and significant impact on public health, economies, and society. COVID-19 is caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and is known for its highly contagious nature and ability to cause severe respiratory infections [1]. The accurate prediction of COVID-19 clinical outcomes and understanding its severity has become essential for effective healthcare resource management. Identifying individuals at higher risk of severe disease is crucial for timely medical interventions, treatment optimization, and improved patient outcomes. Furthermore, predicting COVID-19 severity aids in prioritizing vaccinations and guiding public health measures for high-risk populations.

Machine learning (ML), a branch of artificial intelligence (AI) that relies on data-driven learning models, has found extensive application in healthcare, including risk assessment, large dataset analysis, and result prediction [2]–[5]. Recently, ML has emerged as a valuable tool for predicting COVID-19 severity [6], [7]. ML algorithms analyze diverse data, including clinical parameters, demographics, and biomarkers, to develop predictive models. These models uncover intricate patterns and correlations that might elude human observation, enabling accurate and timely disease severity prediction. While earlier prognosis for COVID-19 patients was limited due to a lack of disease knowledge, ML and AI techniques have enabled accurate outcome prediction. These models focus on self-improvement through learning from examples, obviating the need for explicit programming. Multiple studies have aimed to create models for early-stage infectious disease onset prediction in patients [8]–[11]. Promising results have emerged from ML approaches to COVID-19 severity prediction [12], [13]. By harnessing these techniques, researchers and healthcare professionals can enhance patient triage, resource allocation, and clinical decision-making.

This study seeks to contribute to the growing understanding of COVID-19 severity prediction by comprehensively analyzing biomarkers, clinical parameters, and ML techniques. The structure of this paper is as follows: section 2 provides a literature overview and highlights prior research on COVID-19 severity prediction. Section 3 outlines the study's methodology, encompassing data collection, preprocessing, feature selection, and model development. Section 4 demonstrates model performance, presents results and analysis, and discusses implications and potential clinical applications. Lastly, section 5 concludes by summarizing key findings, addressing limitations, and suggesting avenues for future research.

2. RELATED WORK

COVID-19 severity prediction research has garnered substantial global interest, leading to a multitude of studies exploring diverse approaches and methodologies. A significant focus of these investigations lies in the identification and analysis of clinical parameters as potential predictors of COVID-19 severity. Numerous research studies have been dedicated to the prediction of COVID-19 severity, employing diverse ML algorithms. A comprehensive overview of related studies is presented in Table 1, based on three key selection criteria: i) the primary focus of the study is on COVID severity prediction; ii) the utilization of ML algorithms plays a pivotal role in the predictive modeling process; and iii) the evaluation and measurement of COVID severity are conducted through the application of predictive models.

Table 1. Literature survey

Ref.	ML models used	SMOTE analysis	Important features	Sample size	Performance
[14]	Multi-layer perceptron (MLP)	No	8	257	MLP: AUC score=0.96, F1-score=0.791, accuracy=0.943, precision=0.848, sensitivity=0.776
[15]	Deep learning	No	11	10937	RF: AUC score=0.869, sensitivity=0.822, accuracy=0.807, specificity=0.787
[16]	Random forest (RF), support vector machine (SVM), logistic regression (LR)	Yes	5	224	RF: AUC score=0.86, sensitivity=0.80, accuracy=0.80, specificity=0.81
[17]	RF, naïve bayes (NB), SVM, k-nearest neighbors (KNN), LR, artificial neural network (ANN)	No	5	422	GNB: AUC score=0.89, sensitivity=83.8%, accuracy=81.2%, specificity=81.2%
[18]	MLP, radial basis function (RBF), general regression neural network (GRNN), SVM, RF	No	32	80	RF: sensitivity=0.6785, accuracy=0.9083, precision=0.9083, specificity=0.978, F1-score=0.7756
[19]	XGBoost (XGB), LR	No	15	3028	XGBoost: AUC score=0.8517, sensitivity=0.7747, accuracy=0.7682, precision=0.7652, F1-score=0.7697
[20]	LR, linear discriminant analysis (LDA), KNN, classification and regression trees (CART), NB, SVM, RF	No	11	992	SVM: sensitivity=0.69, accuracy=0.6, precision=0.95, F1-score=0.8
[21]	LR, XGB, RF	Yes	7	287	RF: sensitivity=0.949, accuracy=0.952, specificity=0.956, F1-score=0.955
[22]	RF	Yes	10	5059	RF: AUC score=0.98
[23]	RF, NB, and gradient boosting	No	7	478	RF: sensitivity=98.6, accuracy=78.4, precision=91, specificity=94.7, F1-score=95

The experimental setup involved the utilization of both base and ensemble models, with and without synthetic minority oversampling technique (SMOTE) analysis, considering datasets with diverse sample sizes. The evaluation of these models was based on the receiver operating characteristic area under the curve (ROC-AUC) score, sensitivity, accuracy, precision, and F1-score. Despite advancements, several challenges and limitations have persisted. One prominent limitation is the relatively small dataset, as highlighted in previous studies [14], [17], [18], [20], [21]. This limitation indicates that existing studies may have been conducted on limited data, potentially leading to biased or less generalizable findings. Another noteworthy limitation is the need to explore additional biomarkers and clinical features. This limitation indicates that researchers should explore and investigate other potential biomarkers and clinical features that may significantly contribute to the predictive power of the model [14]. Furthermore, addressing the issue of imbalanced datasets, and employing appropriate feature selection techniques is a challenge in ensuring reliable predictions. The need to discern the most pivotal features and optimal ML techniques is crucial for precise COVID-19 severity prediction.

3. METHODOLOGY

3.1. Data

In this study, we used a dataset obtained from the public dataset by Bhat *et al.* [14]. The dataset was originally collected from 257 confirmed COVID-19 patients admitted to the DY Patil group of hospitals in Pune, Maharashtra, India between July and September 2020. The confirmation of COVID-19 cases was based on positive real-time polymerase chain reaction (RT-PCR) tests for SARS-CoV-2 infection. Patient records were collected and anonymized at the Council of Scientific and Industrial Research (CSIR)-Institute of Genomics and Integrative Biology (CSIR-IGIB) data warehouse to ensure data privacy and confidentiality. This dataset contains 50 features, including demographic and clinical information such as oxygen saturation, respiratory rate, body mass index (BMI), age, sex, comorbidities, and respiratory support levels. In addition, detailed blood test reports were included for 31 different test parameters, including C-reactive protein (CRP), interleukin 6 (IL-6), total leukocyte count (TLC), D-dimer, and lactate dehydrogenase.

3.2. Data pre-processing

Before conducting our analysis, we performed additional data preprocessing to address any potential data quality issues and standardize the data for further analysis. The data were initially examined to identify and remove irrelevant columns such as patient id, area, date of collection, etc. resulting in the removal of seven columns. To address missing data, another seven columns with a higher percentage of missing values were dropped from the dataset because imputing these missing values could introduce bias. Additionally, two columns containing constant values and zeros were identified and removed from the dataset as they did not offer any meaningful information for analysis. To enrich the dataset with valuable information, feature extraction was performed on the symptoms presented and co-morbid condition columns, resulting in the creation of 14 new columns. Moreover, a new target column, severity, was generated by combining the information from the existing Outcome and ventilatory support required columns. If a patient required ventilator support or demise, the corresponding entry in the severity column was marked as severe else no severe, facilitating the classification task [14]. Because data quality and completeness are critical for accurate analysis, all rows with missing values, particularly for biological features, were excluded from the dataset. Consequently, the dataset was refined and 151 records with 45 relevant features were retained for further analysis.

Figure 1 illustrates the distribution of no-severe and severe COVID-19 cases categorized by sex. The data highlight a higher number of male patients affected by the virus, as both no-severe and severe cases are more prevalent among males.The age distribution histogram in Figure 2 provides valuable insights into the relationship between age and the severity of COVID-19 symptoms. The majority of COVID-19 patients in the dataset fall within the age range of 30-70, with a higher number of patients displaying no-severe symptom compared to severe symptom within this age bracket. However, beyond the age of 70 years, there was a significant increase in patients experiencing severe symptoms, with a notable peak observed. Moreover, patients over the age of 80 are more prone to experiencing severe symptoms than no-severe. These observations underscore the critical role of age in determining the severity of COVID-19 symptoms and have important implications for healthcare policies aimed at addressing the prevention and treatment of COVID-19. Upon inspecting the target variable severity, it was evident (Figure 3) that the dataset exhibited an imbalanced class distribution, with an approximate ratio of 1:6 between the severe and no-severe classes.

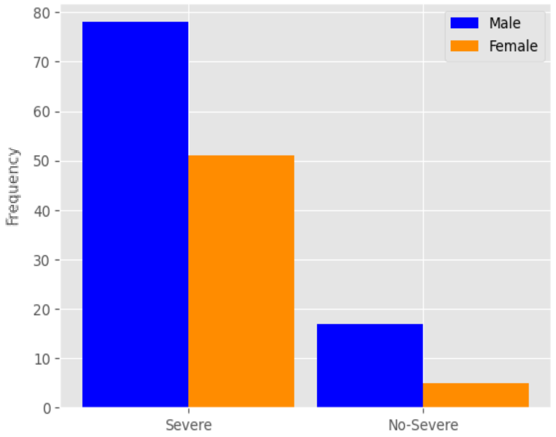

Figure 1. Distribution of patients with severe and non-severe COVID-19 based on sex

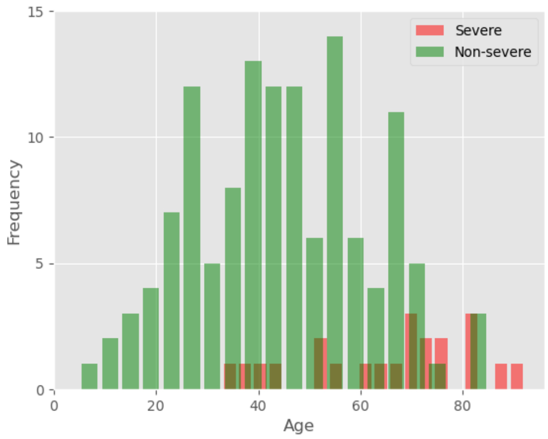

Figure 2. Distribution of severe and no-severe COVID-19 patients based on age

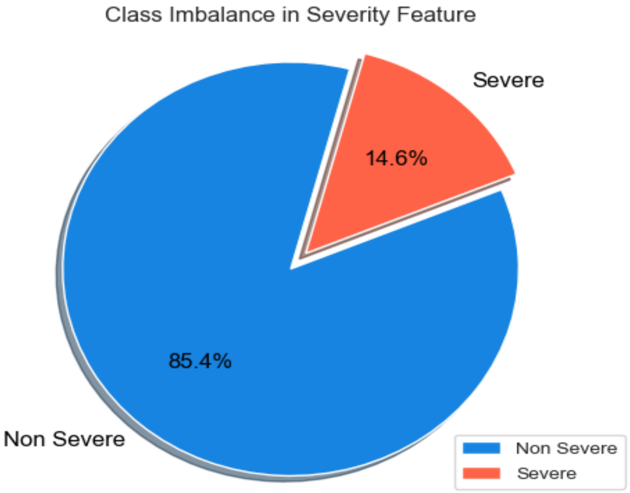

Figure 3. Distribution of severe and non-severe patients

3.3. Proposed model

In this study, we embarked on feature-engineering techniques to identify the most relevant and informative features. We employed various methods, including f-test analysis of variance (ANOVA), mutual information (MI), uniform manifold approximation and projection (UMAP), and principal component analysis (PCA), to extract salient information from the dataset. Each technique yielded a reduced dataset with selected features, which were then split into training and testing sets at a 70:30 ratio. To identify the best feature engineering model, we employed conventional ML techniques, namely, SVM, LR, decision tree (DT), KNN, and Gaussian Naive Bayes (GNB), on each of the reduced datasets. This step allowed us to assess the performance of each technique and to identify the most effective feature selection approach. The selection of the LR, DT, KNN, SVM, and GNB ML algorithms for our research on COVID-19 severity prediction is grounded in robust reasoning and notable attributes: i) proven success in health disorder diagnosis and treatment [6], ii) flexibility in handling complex classification tasks [8], and iii) prominence in the ML community [6]. Leveraging ensemble learning methods, including RF, XGB, AdaBoost, ExtraTrees, voting ensemble, enhanced voting ensemble [5], and stack-based algorithm with meta-learners SVM (SBA_meta_SVM) and GNB (SBA_meta_GNB), we harnessed the collective power of multiple base models to achieve accurate and robust predictions. Figure 4 illustrates the comprehensive workflow of the entire process undertaken in this study.

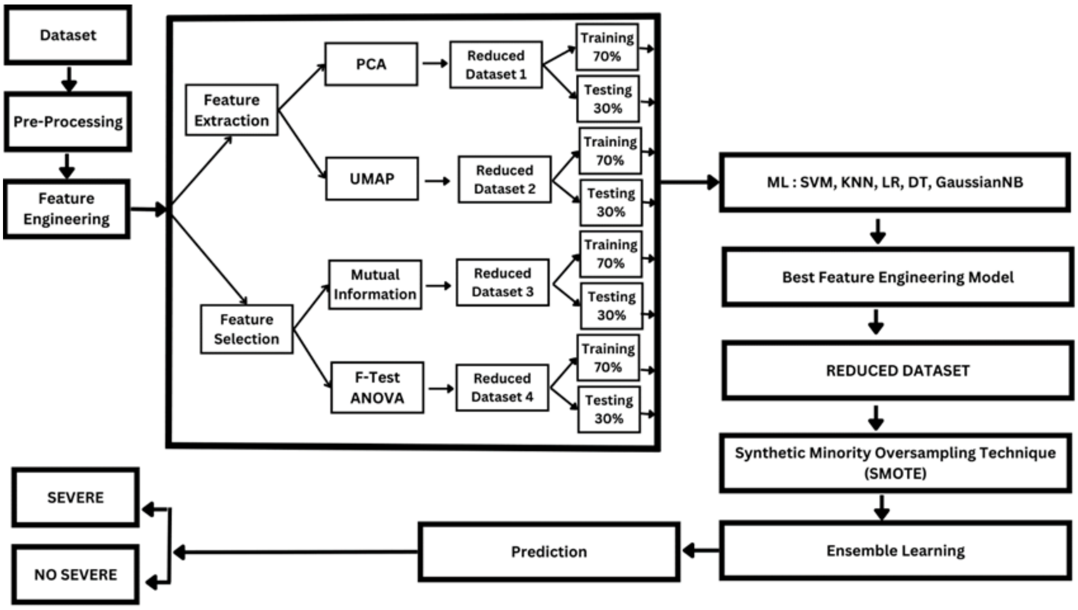

Figure 4. Workflow of the proposed model

3.4. Feature engineering techniques

In the context of ML classification problems, dimensionality reduction and feature selection serve as solutions to overcome the challenges associated with a large number of features. High dimensionality poses difficulties in visualization, analysis, overfitting, and accuracy. The following feature engineering techniques were applied in this study.

- F-Test ANOVA is a statistical method that compares the means of two or more groups to determine if there is a significant difference between them. It assumes normal distribution and equal variances and calculates the ratio of variance between groups to variance within groups to determine the statistical significance of the difference between means. It can be used in various applications, such as clinical trials or evaluating the performance of different groups.
- MI is a measure of the amount of information that two variables share. It is calculated by comparing the joint distribution of the variables to their individual distributions. It is often used in feature selection and dimensionality reduction to identify the most informative features for a given task.
- PCA is a powerful technique for dimensionality reduction that can be used for feature selection in ML. By identifying the most important features and reducing the dimensionality of the dataset, it can improve the performance and stability of ML models.
- UMAP is a nonlinear dimensionality reduction technique used to visualize high-dimensional data in a lower-dimensional space. It works by constructing a topological representation of the data and preserving the local structure of the data in the lower-dimensional space. UMAP has become increasingly popular due to its ability to better preserve the global and local structure of the data compared to other dimensionality reduction techniques.

3.5. SMOTE

SMOTE [24], [25] is an oversampling method that addresses class imbalance by generating synthetic examples for the minority class. SMOTE selects k nearest neighbors from the minority class for each instance and constructs new synthetic instances by interpolating between the feature values of the selected instance(s) and those of its nearest neighbors. The new instances are located in the same region of the feature space as the original instances but differ slightly in their feature values. SMOTE repeats this process until the desired level of oversampling is achieved.

3.6. GidSearchCV

In this study, hyperparameter tuning was performed using GridSearchCV(), a widely used technique for optimizing ML models. This method systematically explores the hyperparameter space through an exhaustive search over a predefined set of hyperparameter values. The model was trained and evaluated for all possible combinations of these hyperparameters using cross-validation, ensuring robust performance evaluation. The optimal hyperparameters were selected based on the maximization of the specified evaluation metric. Adoption of GridSearchCV() in this study facilitated the identification of the most suitable hyperparameter settings for our model, enhancing its performance on unseen data.

3.7. Machine learning techniques

In this research, the following ML techniques have been employed:

- LR is an algorithm in supervised ML for binary classification problems. The algorithm predicts the output of a categorical dependent variable by considering a given set of independent variables. This approach anticipates the outcome of binary events, such as yes or no.
- SVM procedure categorizes linear and non-linear data. SVM uses non-linear mapping to transform the training set to a high level. In this new dimension, SVM searches for the optimal linear hyperplane to separate tuples of one class from another as a decision boundary. A hyperplane with the appropriate non-linear mapping in higher dimensions can separate two-class data. In contrast to the other approaches, hyperplanes are robust for overfitting.
- KNN is a non-parametric algorithm that classifies data points based on their neighbors. It assigns a class label by considering the majority vote of the k nearest neighbors. It is simple but can be computationally expensive for large datasets.
- DT is a tree-based algorithm that splits data based on features to create decision rules. It recursively partitions data until reaching homogeneous subsets. It is interpretable, handles both numerical and categorical data, and can be prone to overfitting.
- GNB is a probabilistic classifier that uses Bayes theorem and assumes a Gaussian distribution of features given class labels. It is computationally efficient and well-suited for high-dimensional datasets with continuous features but may not perform well when assumptions of normality and independence are not met. GNB is widely used in ML and serves as a baseline for comparing more complex algorithms.
- XGBoost is an ensemble-based technique [7] for classification and regression that is a regularized form of the gradient-boosting algorithm. One issue with gradient boosting algorithms is the potential for model overfitting due to data imbalance. XGB addresses this by incorporating a regularization parameter that reduces the risk of overfitting. XGB is also a tree-based ensemble classifier, and it uses a boosting data resampling method to minimize the misclassification error and improve accuracy. This method is iterative and utilizes records that were not successfully predicted in previous iterations for training in subsequent iterations until an optimal result is achieved.
- AdaBoost is a ML algorithm used for classification tasks. It combines multiple weak classifiers to create a strong classifier. It updates sample weights based on weak classifier performance and aggregates their predictions for the final result. AdaBoost is popular for its ability to enhance classification model accuracy and has found wide application in ML.
- RF is an ensemble classifier that uses decision-making with various types of trees. It evaluates the division to create a DT using an arbitrary sequence of features at each node, and each tree is based on the individual values of a random variable. To increase the trees, we can use bagging along with the selection of the random attribute using the CART method. RF uses a random linear combination of the input attributes, and new attributes are created, reflecting a linear combination of existing features rather than choosing the sub-cluster of features randomly.
- ExtraTrees ML algorithm utilizes an ensemble of DT to make predictions. It is an extension of the RF algorithm that employs a technique known as extremely randomized trees. In contrast to RF, ExtraTrees chooses split points randomly for each feature and then selects the optimal split among them. This approach increases the robustness of the model to noisy and irrelevant features and reduces the risk of overfitting. ExtraTrees has demonstrated high performance in classification and regression tasks and is particularly well-suited to handling large datasets with high-dimensional features.
- Voting ensemble, also known as majority voting, is a popular ensemble learning method that combines the predictions of multiple individual models to improve the overall accuracy of a ML system. The technique involves training multiple base models on the same dataset using different learning algorithms or hyperparameters and then aggregating their predictions to arrive at a final decision. The ensemble can be configured to use a simple majority vote or to weight the predictions of each model according to their performance.
- Enhanced-weighted voting ensemble [5] is an extension of the traditional voting ensemble method that assigns different weights to individual classifiers based on their performance on the training set. Here, the algorithm calculates weights and individual weights for a set of inputs and then classifies another set of inputs into positive or negative classes based on the sum of weights in each class. The performance of EWE has been evaluated through experiments on different datasets and compared with other ensemble

methods, demonstrating its effectiveness in improving classification accuracy and handling imbalanced datasets.

- Stacked ensemble methods are a type of ML algorithm that combines multiple models to create a more accurate and robust model. In this approach, several base models are trained independently, and their predictions are combined to create a meta-model. The meta-model is trained on the predictions made by the base models and aims to learn the optimal combination of the base models' predictions to make a final prediction. Stacked ensemble methods are often used in classification and regression problems and can improve the accuracy of predictions by reducing bias and variance. However, they can also be computationally intensive and require careful tuning of hyperparameters.

3.8. Model evaluation

The performance of the model was meticulously assessed using a range of key evaluation metrics: accuracy, which measures overall correctness; precision, indicating the proportion of true positives among predicted positives; sensitivity, gauging the true positive rate; specificity, representing the true negative rate; ROC-AUC, which evaluates the model's ability to distinguish between classes; and the F1-score, which balances precision and sensitivity.

4. RESULTS AND DISCUSSION

The results of our study demonstrate the effectiveness of various feature engineering techniques in reducing the dimensionality of the pre-processed dataset for COVID-19 severity prediction. Applying the F-Test ANOVA led to a reduced dataset with only 16 features, selected based on significant p-values (<0.05). Similarly, mutual information yielded reduced datasets with 10 features, selected based on mutual information scores. UMAP, and PCA yielded reduced datasets with 5 and 13 optimal value of n_components respectively. Subsequently, conventional ML techniques, including SVM, KNN, LR, DT, and GNB, were applied to each reduced dataset derived from the feature engineering techniques. A comprehensive comparison of the results for different feature engineering techniques is presented in Table 2, and Table 3 displays the weighted features obtained using these techniques.

Remarkably, the DT model on the reduced dataset obtained by F-Test ANOVA, emerged as the best feature engineering model, showing superior predictive performance. This model, consisting of only four features, namely, respiratory rate (breaths per minute), C-REACTIVE PROTEINS, age, and TLC COUNT proved to be highly effective in predicting COVID-19 severity, as shown in Figure 5.

Table 2. Comparison of results for different feature engineering techniques

Feature engineering	Model	Accuracy	F1-score	ROC-AUC
MI	KNN	0.804	-	0.474
	GNB	0.804	0.571	0.827
	DT	0.87	0.4	0.631
	SVM	0.827	0.429	0.663
	LR	0.76	0.153	0.507
ANOVA	KNN	0.804	-	0.474
	GNB	0.76	0.476	0.741
	DT	0.891	0.545	0.701
	SVM	0.869	0.625	0.805
	LR	0.847	0.533	0.734
PCA	DT	0.804	0	0.474
	DT	0.804	0.571	0.826
	GNB	0.869	0.4	0.63
	DT	0.826	0.429	0.663
	GNB	0.76	0.154	0.507
UMAP	GNB	0.76	0.267	0.566
	KNN	0.848	-	0.5
	DT	0.848	0.364	0.618
	SVM	0.717	0.134	0.482
	LR	0.76	0.154	0.508

Table 3. Weighted features through feature engineering

Sets	Features selected
Mutual information set	Respiratory rate (breaths per minute), C-REACTIVE PROTEINS, age, Urea
F-Test ANOVA set	Respiratory rate (breaths per minute), C-REACTIVE PROTEINS, age, TLC COUNT
PCA set	13-components
UMAP set	5-components

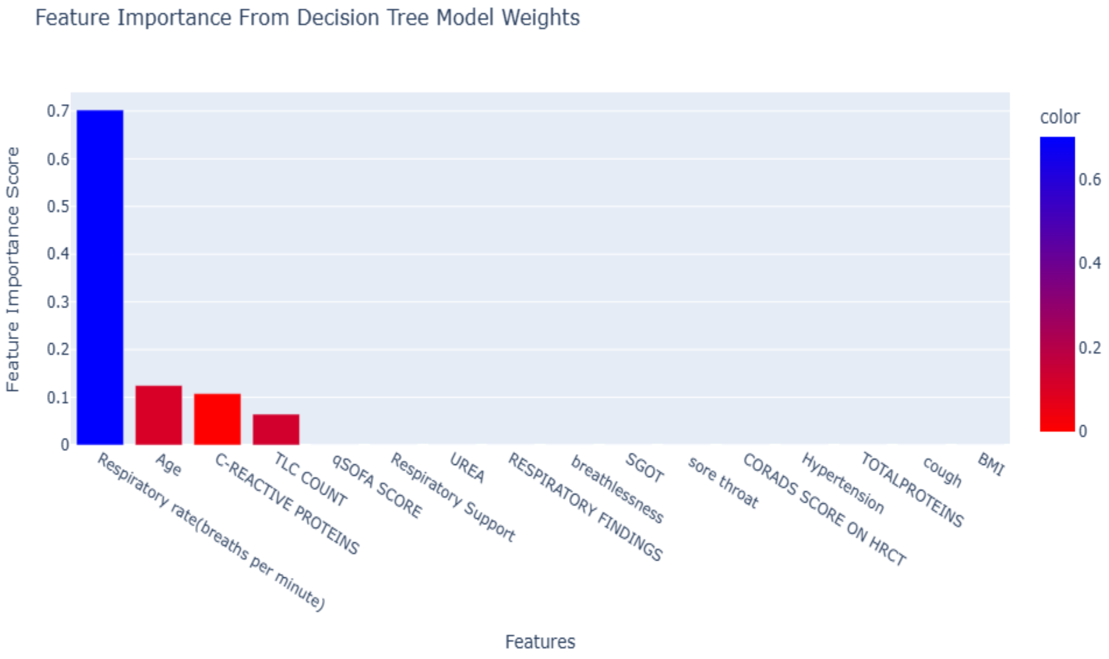

Figure 5. Feature importance scores of the best feature engineering model (DT)

To enhance our predictive models further, various ensemble techniques were employed on the final reduced dataset. Among them, voting ensemble achieved the highest performance, with an ROC-AUC, accuracy, sensitivity, specificity, precision, and F1-score of 0.821, 0.934, 0.714, 0.974, 0.834, and 0.769 respectively, as presented in Table 4.

Table 4. Comparison of ensemble models performance-without SMOTE

Ensemble models	ROC-AUC	Accuracy	Sensitivity	Specificity	Precision	F1-score	Time elapsed (s)
XGBoost	0.86	0.934	0.571	1	1	0.727	6.206
RF	0.831	0.913	0.571	0.974	0.8	0.667	114.583
AdaBoost	0.714	0.913	0.571	0.974	0.8	0.667	0.110
ExtraTrees	0.9	0.934	0.571	1	1	0.727	100.672
Voting ensembles	**0.821**	**0.934**	**0.714**	**0.974**	**0.834**	**0.769**	**0.063**
Enhanced weighted voting ensemble	0.812	0.934	0.571	1	1	0.727	0.047
SBA-GNB	0.667	0.891	0.714	0.923	0.625	0.667	0.078
SBA-SVM	0.614	0.847	0.571	0.897	0.5	0.534	0.078

The original dataset exhibits class imbalance, which can potentially impact the model's performance. To overcome this issue, SMOTE is applied to the training set, resulting in a balanced dataset that ensures both classes are equally represented. Prior to SMOTE, the training set comprised 15 severe cases and 90 non-severe cases, highlighting the severe class imbalance. However, after applying SMOTE, the number of samples in the training set increased to 180, with an equal representation of 90 samples for both the severe and non-severe classes. The application of SMOTE to address class imbalance resulted in an even more robust performance. The AdaBoost emerged as the top-performing ensemble model, showing remarkable ROC-AUC, accuracy, sensitivity, specificity, precision, and F1-score of 0.965, 0.978, 1.0, 0.974, 0.875, and 0.934 respectively, with a swift execution time of 0.088 s, as shown in Table 5.

Figure 6 provides a visual representation of ROC curves of different ensemble models. Figure 7 presents a comprehensive comparison of the evaluation parameters for different models using SMOTE. The evaluation metrics used in the comparison included ROC-AUC, accuracy, sensitivity, specificity, precision, and F1-score.

Table 5. Comparison of ensemble models performance-with SMOTE

Ensemble models	ROC-AUC	Accuracy	Sensitivity	Specificity	Precision	F1-score	Time elapsed (s)
XGBoost	0.998	0.935	1.0	0.923	0.7	0.823	6.917
RF	0.999	0.935	0.857	0.948	0.7	0.800	121.372
AdaBoost	**0.965**	**0.978**	**1.0**	**0.974**	**0.875**	**0.934**	**0.0888**
ExtraTrees	0.999	0.913	0.857	0.923	0.667	0.750	117.986
Voting ensembles	0.994	0.870	0.857	0.871	0.545	0.667	54.419
Enhanced weighted voting ensemble	0.996	0.870	0.857	0.871	0.545	0.667	57.457
SBA-GNB	0.977	0.913	1.0	0.897	0.636	0.778	49.572
SBA-SVM	0.937	0.652	0.28	0.71	0.153	0.200	46.923

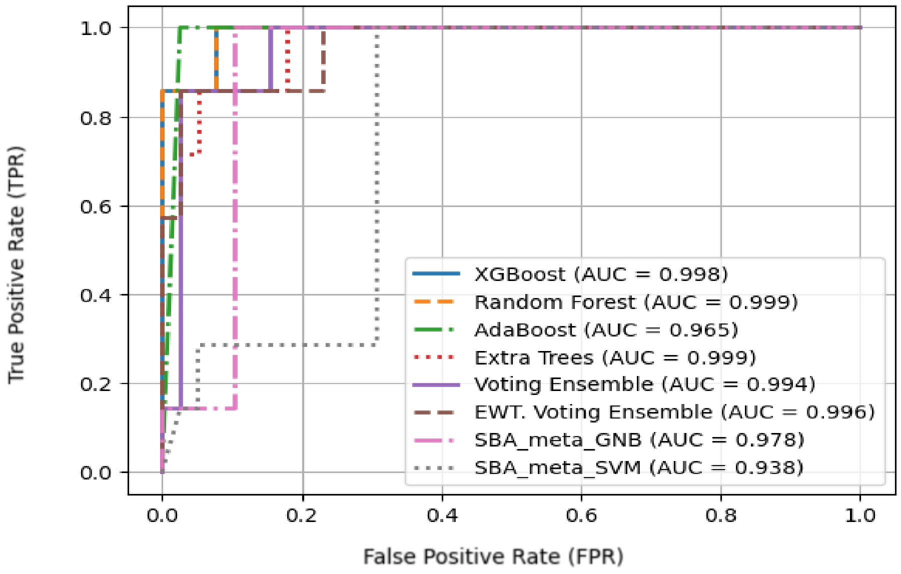

Figure 6. ROC-AUC curve for different ensemble models

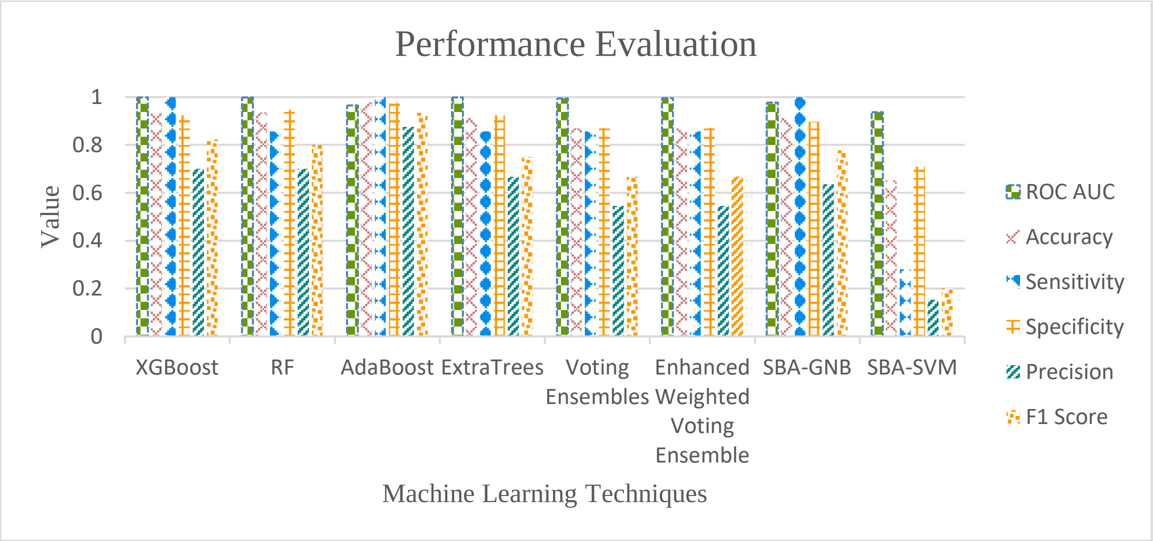

Figure 7. Comparison of performance evaluations (with SMOTE) for different ML models

In this study, we effectively addressed several challenges and limitations identified in the literature survey, thereby enhancing the robustness and applicability of COVID-19 severity prediction models. Our research presents significant achievements in overcoming these challenges, contributing to the advancement of predictive modelling for COVID-19 severity assessment.

To address the limitations of a limited dataset, we employed various feature engineering techniques, including F-Test ANOVA, mutual information, UMAP, and PCA. These techniques effectively reduce the dimensionality of the pre-processed dataset, allowing us to retain essential features while mitigating the potential loss of information due to random undersampling. Furthermore, by employing conventional ML techniques and ensemble methods on a reduced dataset, scalable and robust prediction models were successfully developed. Our top-performing ensemble model, the Adaboost, demonstrated exceptional accuracy, F1-score, sensitivity, and ROC-AUC metrics, meeting the need for better predictive performance.

We have also addressed the issue of data imbalance by using the SMOTE. Imbalanced datasets are common in clinical settings, where the number of severe COVID-19 cases is often significantly lower than that of non severe cases. The incorporation of SMOTE in our predictive modelling approach not only addresses the data imbalance challenge, but also contributes to the overall robustness and generalizability of the COVID-19 severity prediction models. This ensures that the models are better equipped to handle real-world scenarios with imbalanced datasets and enhances the model's reliability in clinical decision-making processes.

Bhat *et al.* [14] reported that the MLP model achieved the highest accuracy of 0.942, an F1-score of 0.791, a precision of 0.847, a sensitivity of 0.776, and a ROC-AUC score of 0.961. However, in our study with the same dataset, the proposed model, exhibits superior performance in terms of accuracy of 0.978, F1-score of 0.934, precision of 0.875, sensitivity of 1.0, and ROC-AUC score of 0.965 making it a promising choice for COVID-19 severity prediction with a substantial improvement in F1-score from 0.791 to 0.934.

5. CONCLUSION

This research highlights the significance of robust feature engineering and ensemble techniques for predicting COVID-19 severity. The DT model derived from the F-Test ANOVA demonstrated superior performance, and the application of SMOTE further enhanced the accuracy and predictive power of the ensemble models. The proposed model outperformed the top-performing models reported in previous studies by achieving an accuracy of 0.978, sensitivity of 1.0, precision of 0.875, F1-score of 0.934, and ROC-AUC of 0.987. These results have promising implications for improving clinical decision-making and patient care in the management of COVID-19 cases.

Looking towards the future scope, we propose the acquisition of larger and diverse datasets from multiple healthcare centers for external validation to enhance the generalizability of predictive models. In addition, we advocate the integration of multiple data sources, such as genomic and imaging data, to capture a comprehensive view of the complexity of the disease. Moreover, we encourage further exploration of explainable AI techniques to improve the interpretability of models and foster better acceptance and trust in the healthcare community.

REFERENCES

[1] K. G. Andersen, A. Rambaut, W. I. Lipkin, E. C. Holmes, and R. F. Garry, “The proximal origin of SARS-CoV-2,” *Nature Medicine*, vol. 26, no. 4, pp. 450–452, Apr. 2020, doi: 10.1038/s41591-020-0820-9.

[2] A. Ouhmida, A. Raihani, B. Cherradi, and S. Sandabad, “Parkinson’s diagnosis hybrid system based on deep learning classification with imbalanced dataset,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, no. 3, pp. 3204–3216, Jun. 2023, doi: 10.11591/ijece.v13i3.pp3204-3216.

[3] S. Krishnan, P. Magalingam, and R. Ibrahim, “Hybrid deep learning model using recurrent neural network and gated recurrent unit for heart disease prediction,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 6, pp. 5467–5476, Dec. 2021, doi: 10.11591/ijece.v11i6.pp5467-5476.

[4] N. Razali, S. Ismail, and A. Mustapha, “Machine learning approach for flood risks prediction,” *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 9, no. 1, pp. 73–80, Mar. 2020, doi: 10.11591/ijai.v9.i1.pp73-80.

[5] V. C. Osamor and A. F. Okezie, “Enhancing the weighted voting ensemble algorithm for tuberculosis predictive diagnosis,” *Scientific Reports*, vol. 11, no. 1, p. 14806, Jul. 2021, doi: 10.1038/s41598-021-94347-6.

[6] S. Mishra, R. Kumar, S. K. Tiwari, and P. Ranjan, “Machine learning approaches in the diagnosis of infectious diseases: a review,” *Bulletin of Electrical Engineering and Informatics*, vol. 11, no. 6, pp. 3509–3520, Dec. 2022, doi: 10.11591/eei.v11i6.4225.

[7] J. Li *et al.*, “An ensemble prediction model for COVID-19 mortality risk,” *Biology Methods and Protocols*, vol. 7, no. 1, Jan. 2022, doi: 10.1093/biomethods/bpac029.

[8] S. Maheshwari, A. Sharma, R. Kumar, and Pratyush, “Early Detection of Influenza Using Machine Learning Techniques,” in *Recent Innovations in Computing*, Singapore: Springer, 2022, pp. 111–124, doi: 10.1007/978-981-16-8892-8_9.

[9] R. Kumar, S. Maheshwari, A. Sharma, S. Linda, S. Kumar, and I. Chatterjee, “Ensemble learning-based early detection of influenza disease,” *Multimedia Tools and Applications*, May 2023, doi: 10.1007/s11042-023-15848-2.

[10] H. Abu Owida, H. S. Migdadi, O. S. Mohamed Hemied, N. F. Fankur Alshdaifat, S. F. Ahmad Abuowaida, and R. S. Alkhawaldeh, “Deep learning algorithms to improve COVID-19 classification based on CT images,” *Bulletin of Electrical Engineering and Informatics*, vol. 11, no. 5, pp. 2876–2885, Oct. 2022, doi: 10.11591/eei.v11i5.3802.

[11] H. Imaduddin, F. Yusfida Ala, A. Fatmawati, and B. A. Hermansyah, “Comparison of transfer learning method for COVID-19

detection using convolution neural network," *Bulletin of Electrical Engineering and Informatics*, vol. 11, no. 2, pp. 1091–1099, Apr. 2022, doi: 10.11591/eei.v11i2.3525.

[12] A. H. Ahmed, M. N. A. Al-Hamadani, and I. A. Satam, "Prediction of COVID-19 disease severity using machine learning techniques," *Bulletin of Electrical Engineering and Informatics*, vol. 11, no. 2, pp. 1069–1074, Apr. 2022, doi: 10.11591/eei.v11i2.3272.

[13] S. Solayman, S. A. Aumi, C. S. Mery, M. Mubassir, and R. Khan, "Automatic COVID-19 prediction using explainable machine learning techniques," *International Journal of Cognitive Computing in Engineering*, vol. 4, pp. 36–46, Jun. 2023, doi: 10.1016/j.ijcce.2023.01.003.

[14] S. Bhat *et al.*, "Learning From Biological and Computational Machines: Importance of SARS-CoV-2 Genomic Surveillance, Mutations and Risk Stratification," *Frontiers in Cellular and Infection Microbiology*, vol. 11, Dec. 2021, doi: 10.3389/fcimb.2021.783961.

[15] V. Singh *et al.*, "A deep learning approach for predicting severity of COVID-19 patients using a parsimonious set of laboratory markers," *iScience*, vol. 24, no. 12, p. 103523, Dec. 2021, doi: 10.1016/j.isci.2021.103523.

[16] B. K. Patterson *et al.*, "Immune-Based Prediction of COVID-19 Severity and Chronicity Decoded Using Machine Learning," *Frontiers in Immunology*, vol. 12, Jun. 2021, doi: 10.3389/fimmu.2021.700782.

[17] R. Zhang, Q. Xiao, S. Zhu, H. Lin, and M. Tang, "Using different machine learning models to classify patients into mild and severe cases of COVID-19 based on multivariate blood testing," *Journal of Medical Virology*, vol. 94, no. 1, pp. 357–365, Jan. 2022, doi: 10.1002/jmv.27352.

[18] A. Alotaibi, M. Shiblee, and A. Alshahrani, "Prediction of Severity of COVID-19-Infected Patients Using Machine Learning Techniques," *Computers*, vol. 10, no. 3, p. 31, Mar. 2021, doi: 10.3390/computers10030031.

[19] L. Jia *et al.*, "An interpretable machine learning model based on a quick pre-screening system enables accurate deterioration risk prediction for COVID-19," *Scientific Reports*, vol. 11, no. 1, p. 23127, Nov. 2021, doi: 10.1038/s41598-021-02370-4.

[20] H. Gull, G. Krishna, M. I. Aldossary, and S. Z. Iqbal, "Severity Prediction of COVID-19 Patients Using Machine Learning Classification Algorithms: A Case Study of Small City in Pakistan with Minimal Health Facility," in *2020 IEEE 6th International Conference on Computer and Communications (ICCC)*, Dec. 2020, pp. 1537–1541, doi: 10.1109/ICCC51575.2020.9344984.

[21] S. S. Aljameel, I. U. Khan, N. Aslam, M. Aljabri, and E. S. Alsulmi, "Machine Learning-Based Model to Predict the Disease Severity and Outcome in COVID-19 Patients," *Scientific Programming*, vol. 2021, pp. 1–10, Apr. 2021, doi: 10.1155/2021/5587188.

[22] T. W. Tulu *et al.*, "Machine learning-based prediction of COVID-19 mortality using immunological and metabolic biomarkers," *BMC Digital Health*, vol. 1, no. 1, p. 6, Feb. 2023, doi: 10.1186/s44247-022-00001-0.

[23] A. M. I. A. Yusuf, M. M. Rosli, and N. S. M. Yusop, "A Screening System for COVID-19 Severity using Machine Learning," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 7, 2022, doi: 10.14569/IJACSA.2022.0130746.

[24] 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, Jun. 2002, doi: 10.1613/jair.953.

[25] S. Uyun and E. Sulistyowati, "Feature selection for multiple water quality status: Integrated bootstrapping and SMOTE approach in imbalance classes," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 4, pp. 4331–4339, Aug. 2020, doi: 10.11591/ijece.v10i4.pp4331-4339.

BIOGRAPHIES OF AUTHORS

Smriti Mishra is pursuing Ph.D. in Computer Science from Magadh University, Bodhgaya, India. She obtained Master of Computer Application (MCA) degree from Jawaharlal Technological University, India 2011. In 2008, she completed her Bachelor's degree in Computer Application (BCA) from Magadh University, Bodh Gaya. Her research interest includes data science, machine learning, image processing, and artificial intelligence. She can be contacted at email: smritiit.gaya@gmail.com.

Ranjan Kumar is an Associate Professor at Department of Computer Science, Aryabhatta College, University of Delhi, Delhi, India. He received his Master degree (MCA) from Department of Computer Science, University of Delhi in 2003 and completed his Ph.D. in Software Reliability in the year 2019. His research interests include software reliability, machine learning, data science and bio medical computation. He can be contacted at email: ranjan301@gmail.com.

Sanjay Kumar Tiwari is currently working as Associate Professor at P.G. Department of Mathematics, Bodh Gaya, India. He has received his Ph.D. from Magadh University, Bodh Gaya, India. He has more than 20 years of academic experience. His area of interest includes fixed-point theorem, real analysis, software engineering, and data science. He is contributing as a reviewer of various reputed journals. He can be contacted at email: tiwari.dr.sanjay@gmail.com.

Priya Ranjan graduated from IIT Kharagpur (EE, 1997), India and earned M.S. (EE) and Ph.D. (ECE) degrees from the University of Maryland, College Park, USA (Est.: 1856) in 1999 and 2003, respectively. To pursue his passion in discovering new ideas in control theory and applications, he joined the Center for Artificial Intelligence and Robotics (CAIR, Bangalore), a DRDO lab under the tutelage of Prof. M. Vidyasagar. To continue and expand his studies, in 1997 he moved to the Computer Aided Control Systems Engineering (CACSE) lab at the Institute for Systems Research (ISR) at the University of Maryland under the guidance of Prof. E.H. Abed where he would spend the next twelve years in Maryland in various capacities. His latest interest is in biological computation and oncological image processing. He has received many large project awards from some of the most reputed agencies like National Science Foundation (NSF)-USA, DARPA (credited with leading development of the internet)-USA, Delhi University, DBT, DST-RFBR, ICMR etc. He has authored more than hundred research articles in different international journals and conferences around the world, many heavily cited journal papers and many book chapters in heavily cited books published by IEEE and Springer Press. He can be contacted at: pranjan@gmail.com.