

COST-SENSITIVE HETEROGENEOUS STACKED ENSEMBLE CLASSIFIER MODEL FOR PREDICTING SARS-COV-2 INFECTION FROM CLINICAL AND LABORATORY DIAGNOSTIC RESULTS

Dhivya P¹, Durgadevi Velusamy^{2*}, Karthikeyan Ramasamy³

¹Information Technology, Sri Krishna College of Engineering and Technology, Kuniyamuthur, Coimbatore, 641 008, Tamilnadu, India.

^{2*}Information Technology, SSN School of Engineering, Shiv Nadar University, Chennai, Kalavakkam, Chennai, 603 110, Tamilnadu, India.

³Electrical and Electronics Engineering, M.Kumarasamy College of Engineering, Thalavapalayam, Karur, 639 113, Tamilnadu, India.

Corresponding author(s): DHIVYA P
dhivyainfo@gmail.com

Abstract: The novel coronavirus disease (COVID-19) was initially identified in Wuhan, China, in December 2019 and subsequently had a profound global impact due to its rapid spread. Earlier symptoms and conditions of this deadly virus share common characteristics with the common cold and influenza, making the diagnosis difficult for health care professionals. This study aims to provide a decision on earlier and reliable screening of COVID-19 patients from the clinical blood tests and laboratory reports. We have proposed a cost-sensitive heterogeneous stacked ensemble classifier (CS-HSEC) that combines three different classifiers, namely random forest (RF), gradient boosting machine (GBM) and K-nearest neighbour (K-NN) at level-1 and naive bayes classifier (NB) as the meta classifier at level-2 for diagnosing COVID-19 disease. We have selected highly significant features using the hybrid feature selection algorithms based on mutual information (MI) technique to enhance the performance of the classifiers. The classifiers are trained and tested on the dataset acquired from the Israelita Albert Einstein Hospital, Sao Paulo, Brazil. The proposed CS-HSEC performance is evaluated using 10 clinical features of complete blood test with stratified 10-fold cross-validation and stratified train-test split, with 90% data for training and remaining for testing. The experimental results prove the proposed model's promising performance, having an accuracy of 98.33% with sensitivity, specificity and AUC of 0.93, 0.99 and 0.96, respectively. The proposed CS-HSEC can potentially find the COVID-19 disease during the initial screening of a patient and can be used as an assisting tool by clinicians in diagnosing the COVID-19 disease.

Keywords: SARS-CoV-2 virus, COVID-19 disease, Mutual Information, Ensemble Learning, Cost-sensitive learning, Laboratory findings.

1. INTRODUCTION:

Coronavirus disease (COVID-19) is an infectious disease caused by the SARS-CoV-2 virus has no respite in spreading across the world. It was originally identified in Wuhan, China, in 2019 and became pandemic in 2020 [1]. A COVID-19 infection is an acute respiratory disease caused by a coronavirus in humans that especially has severe symptoms and, in some cases, leads to death in older people and those with health complications like heart disease, kidney failure, diabetes and so on [2]. The symptoms associated with SARS-COV-2 diseases are fever, cough, tiredness, and loss of taste and smell. Based on the severity of the infection, the patients may also have difficulty breathing and chest pain that requires immediate medical attention as the symptoms worsen [5]. In such medical casualties, faster decision-making may reduce the mortality and morbidity rate of COVID-19 patients. The major challenge faced in identifying and classifying COVID-19 infection is the severity of lung infection caused by other diseases like pneumonia and flu.

Medical diagnosis refers to the identification of a disease or condition that accounts for a patient's symptoms and clinical signs. Several testing methods have been developed to diagnose the COVID-19 disease. The COVID-19 disease is medically diagnosed using the standard diagnostic methods like transcription-mediated amplification (TMA), real-time reverse transcription-polymerase chain reaction (rRT-PCR), or reverse transcription loop-mediated isothermal amplification (RT-LAMP) from a swab.

Although RT-PCR identifies COVID-19 infections, the availability of the kit is limited, and it suffers a high false-negative rate with an 80% of accuracy rate [3], [4]. Hence, medical images are used as an alternative for identifying SARS-CoV-2 infections. The SARS-CoV-2 infections can be identified effectively using chest X-rays, CT scans and blood samples to diagnose the severity of the COVID-19 in a person. The most widely used modality for diagnosing COVID-19 disease is a chest X-ray to identify the lung's infection [5], [7]. However, the physical diagnosis of X-ray requires an experienced radiologist and takes a longer diagnostic time for the experts as the number of patients grows.

The identification of COVID-19 disease can be enhanced by using artificial and computational intelligent techniques that help in the effective detection and treatment of the disease. The machine learning algorithm utilizes the available data to learn the underlying relationships between different attributes. Then, the learned model can be used to identify disease in a patient. Deep learning algorithms are the emerging field that shows promising results in identifying COVID-19 infections from chest X-rays, or CT images [7]. Deep neural networks are used in detecting COVID-19 infections that can assist the radiologist in identifying the disease effectively [3], [5]. However, a Deep neural network incurs a high computational cost for the feature extraction and training time for an image dataset. The computational time increases with an increase in data size. Hence, in this study, COVID-19 infection detection is performed using the data acquired from the patients' clinical and laboratory test results. To enhance the performance of the machine learning algorithm, it is essential to identify the relevant features for identifying and classifying COVID-19 infection.

Feature selection plays a vital role in identifying significant features in predicting COVID-19 disease [22]. The selection of appropriate features reduces training cost, data redundancy and helps to improve the predictive performance of the model. The feature importance measures like random forest variable importance [17], [19], glmnet-based feature importance [17], artificial bee colony optimization [29] and so on. In this research, a hybrid combination of mutual information (MI) based feature selection is used for selecting highly significant, and relevant feature attributes for the classification of COVID-19 [23]. The main motivation of this study is to develop a stacked ensemble-classifier model based on cost-sensitive learning with MI based optimal feature subset to enhance the performance of COVID-19 classification by reducing the total misclassification cost.

This study proposes a cost-sensitive heterogeneous stacked ensemble classifier model (CS-HSEC) with weighting-based cost-sensitive learning with MI based feature selection for identifying COVID-19 disease utilizing the publicly available clinical dataset containing information about COVID-19 patients acquired from Hospital Israelita Albert Einstein, Sao Paulo, Brazil [21]. The clinical dataset consists of results of a complete blood test, urine, diabetes, renal function, liver function and adrenal function tests. The proposed heterogeneous stacked ensemble model is constructed with two stacking levels that combine three cost-sensitive classifiers: random forest [6], gradient boosting machine and K-nearest neighbours in the first level and the Naive Bayes classifier as meta-classifier in the second level. The proposed model exploits the structural diversity of the base classifiers, helping to capture each classifier's underlying data distribution for improving the classification of COVID-19 disease. To the author's knowledge, this is the first study that has applied cost-sensitiveness in stacked-ensemble learning along with MI based feature subset selection to diagnose SARS-CoV-2 virus infection using the clinical and laboratory findings.

The classifiers at the first level in the stacking model are trained using training data, and the unseen test data are used to predict the individual algorithms' performance. The predictive performance of the CS-HSEC along with the individual classifiers as assessed in terms of detection accuracy, recall, precision, F measure, kappa, specificity, Matthew's Correlation Coefficient (MCC), and Area under the Curve (AUC) in COVID-19 diagnosis. The test performance validates the efficacy of the CS-HSEC in effectively diagnosing COVID-19 infection with an accuracy of 98.3%, recall of 0.93 and specificity of 0.99, with a precision of 96% in COVID-19 diagnosis. These diagnostic results are comparatively higher than the other state-of-the-art techniques in the literature. The main contribution of this study is as follows:

- We have proposed a cost-sensitive heterogeneous stacked ensemble classifier for identifying SARS-CoV-2 infection from clinical and laboratory test results by forming two-level stacking with 10-fold cross-validation. The stacked-ensemble algorithm combines the structural diversity of the random forest, gradient boosting machine, and K-nearest neighbour at the first level. The

probability prediction of the classifiers in first level is given as the input to the Naive Bayes classifier as a meta-learner at second-level of stacking.

- We have applied a hybrid feature selection technique that combines mutual information (MI) technique to select highly significant and discriminant features for diagnosing COVID-19 disease. We have selected ten highly discriminant features for the COVID-19 classification process.
- To minimize the misclassification cost and handle the highly imbalanced dataset (ratio of 6.5:1), weighting-based cost-sensitive learning is used to compute the normalized weight for each class based on the class distribution.

The remaining sections are organized as follows. A review of the previous literatures related to imbalanced class distribution problem in medical diagnosis with related publication in clinical diagnosis of COVID-19 infection using laboratory findings are presented in Section 2. Section 3 discuss about the normalization methods, Mutual Information-based feature selection algorithm, and proposed cost-sensitive heterogeneous stacked ensemble algorithm used in the detection of COVID-19 disease. Section 4 briefly describes about the performance evaluation metrics used to assess the developed classifiers performance with the experimental results and discussion about the diagnostic results of the proposed CS-HSEC approach on the laboratory and clinical findings dataset of COVID-19 infection. Finally, Section 5 concludes the paper with scope for further research.

2. RELATED WORKS

Artificial intelligence and Machine learning-algorithms has found popularity and have been applied widely in the fields of image processing, heart disease identification [8], cancer diagnosis, alzheimer disease and so on. With the advancement in clinical diagnostic system and its application in medical diagnosis helped the medical experts to conduct a comprehensive study in analyzing the clinical laboratory findings for better understanding of the disease. However, the imbalanced class distribution in medical dataset is a greater challenge for both researchers and medical experts.

2.1 REVIEW OF IMBALANCED CLASS DISTRIBUTION PROBLEM IN MEDICAL DIAGNOSIS

Despite the latest advancements in clinical diagnosis and decision support system (CDSS) for disease diagnosis and prediction using machine learning algorithms, one of the problems in disease diagnosis and classification is the way of handling class imbalance problem [8]. Class imbalance has been extensively investigated, as most machine learning algorithms operate under the assumption of equal sample distributions among all classes within a dataset. However, in real-time applications, the class distribution is uneven with number of samples in majority class (healthy or negative class) is relatively higher than the number of samples in the minority class (diseased or positive class). In literature, several studies deal with the imbalanced class distribution problem using resampling techniques [9] and cost-sensitive learning algorithms. The resampling technique tries to balance the dataset at data-level by applying over-sampling method or under-sampling method and sometimes, applies both techniques for balancing the dataset. Thus, the imbalanced dataset can be balanced to have even distribution of class samples using resampling techniques [10]. Nonetheless, during the resampling process, these resampling techniques change the underlying distribution of classes that may result in omitting some of the important data instances (majority class) or increases the computational cost by adding more instances (minority class) [11].

Cost-sensitive learning addresses imbalanced class distributions by assigning specific costs to the misclassification of data instances. In contrast to resampling technique, cost-sensitive learning works at algorithm-level by incorporating a class weight in the cost-insensitive algorithm making the classifier cost-sensitive. A cost or weight matrix is utilized for assigning cost/weight for misclassification of classes during the training process of machine learning algorithm [11]. The approach aims to minimize misclassification costs, which in turn enhances the model's classification performance on the input data. Cost-sensitive learning is used in many research works for medical data diagnosis that employs machine learning algorithms like artificial neural network, ensemble algorithm [12], XGBoost, support vector machine [11], AdaCost [13], multi-layer perceptron, deep neural network, among others. These

algorithms are employed for applications like diabetes [10], chronic kidney disease [12], heart disease, liver disease, dermatology [13], cancer diagnosis [11], [14], and so on to identify the minority class samples accurately.

2.2 RESEARCH ON CLINICAL DIAGNOSIS OF COVID-19 INFECTION USING LABORATORY FINDINGS

Numerous studies have assessed the performance of machine learning and deep learning algorithms in various medical imaging modalities, including chest X-rays, CT scans, ultrasound, and clinical laboratory datasets, for the accurate detection of COVID-19 infections. Although there is limited literature on clinical prediction of COVID-19 using laboratory parameters, several studies have utilized machine learning algorithms to predict COVID-19 based on laboratory findings.

The authors in [24] applied machine learning algorithm to predict the severity of COVID-19 infection using the clinical data obtained from Wenzhon central hospital and Cangnan peoples hospital in Wenzhon, China. The authors have developed Machine learning algorithms namely random forest (RF), decision tree (DT), logistic regression (LR), K-nearest neighbors(KNN) and support vector machine (SVM) for classification of COVID-19 disease and achieve a classification accuracy of 80% using SVM classifier. An investigation is done in [19] using different machine learning algorithms namely, KNN, Naive Bayes, LR, extremely randomized trees, DT, RF, SVM and three-way RF classifier for prediction of COVID-19. The 3-way RF model was trained and obtained an accuracy of 86%, with 95% sensitivity. The authors in [28] combined the findings of computer tomography with laboratory data to identify the COVID-19 infection using convolution neural network and machine learning and achieved sensitivity of 84.3%, specificity of 82.8% and 92% AUC. The following are the studies in the literature that have used the clinical and laboratory test dataset obtained from Hospital Israelita Albert Einstein at Sao Paulo, Brazil [21] for prediction of COVID-19 disease.

The authors in [15] developed different deep learning models using 10-fold cross validation and train-test split approach and achieved a classification accuracy of 92.9% with sensitivity of 93.68% and precision with 92.35% using convolution neural network with short term memory network. The study in [16] developed an ensemble learning model with RF, GBM and LR as base classifiers and XGBoost as meta classifier with synthetic minority oversampling technique (SMOTE) for COVID-19 prediction. The developed stacked ensemble model attained a classification accuracy of 99.88% with 98.72% recall and 99.99% specificity in predicting the disease.

An initial screening was done based on the blood test to identify COVID-19 patients in [17] using four machine learning algorithms namely RF, LR, ANN and generalized linear model network (GLMNET) with 14 features. The developed models are trained and achieved an accuracy of 87% with 86% Area Under the Curve. A COVID-19 diagnostic system is created and developed as an application in [18] that utilizes multi-layer perceptron (MLP), SVM, Bayesian Network (BN), NB, RF and Randomized trees with SMOTE. The Bayesian Network model has superior performance with overall accuracy of 95.19%, a precision of 93.8%, sensitivity of 96.8% and 93.6% specificity in comparison with other models.

A study in [20] used machine learning algorithm to predict SARS-COV-2 infection. The authors developed five different classifiers namely SVM, Neural Network, Logistic Regression, random forest and gradient boosting machine. The SVM and RF classifier achieved the AUC of 0.847. The authors in [27] developed a clinical predictive model using various machine learning algorithm for classification of COVID-19 disease learning algorithms like RF, NN, LR, SVM and XGB and the extreme gradient boosting achieved an AUC of 66% in identification of the disease. The authors in [29] developed Deep learning classifier with newly created features using pre-weighty vector optimization using artificial bee colony and applied SMOTE preprocessing technique to evaluate the performance of the RNN model. The developed model achieved an accuracy of 94.95% in predicting SARS-CoV-2 virus infection.

Some authors have applied the widely used Synthetic Minority Oversampling Techniques (SMOTE) resampling technique for improving the predictive accuracy of the model by balancing the minority class in an imbalanced dataset [16], [25], [29]. However, SMOTE resampling techniques add the instances into minority class thereby changes the skewed distribution of dataset and increases the computational cost. However, all the above studies have not applied cost-sensitive learning for

identification of SARS-CoV-2 infection. In the present study, we have added class weights with the base classifiers and stacked three cost-sensitive base classifiers namely Random Forest, Gradient Boosting Machine and K-Nearest neighbors with Naive Bayes algorithm as meta-classifier to predict the COVID-19 infection from clinical findings and laboratory data.

3. MATERIALS AND METHODS

This section discuss in detail about the dataset used for the detection of COVID-19 infections. The next subsection describes the feature selection algorithm used for selecting significant features. The data preprocessing technique followed by cost-sensitive classification are described in the consecutive subsection. Subsequently, the proposed cost-sensitive, stacked ensemble algorithm and its implementation is given in the last section.

3.1. MEDICAL DATASET USED

The dataset comprises anonymized records from 5,644 individual patients collected by Hospital Israelita Albert Einstein in Sao Paulo, Brazil, between 28 March to 3 April 2020, and is utilized for the identification of COVID-19 disease [21]. The dataset provides information about the patient admitted to regular ward,

semi-intensive unit and intensive care unit along with their medical diagnostics. The data is collected perform RT-PCR test and other clinical and laboratory tests like complete blood and urine tests with a total of 111 laboratory findings. From the dataset, only 600 sample records are used as the remaining samples lacks laboratory test results. We have used the dataset consists of 600 samples with 18 laboratory findings namely Haematocrit, Red blood Cells, Lymphocytes, Eosinophils, Haemoglobin, Platelets, Monocytes, Leukocytes, Basophils, Serum Glucose, Neutrophils, Creatinine, Urea, Proteina C reativa mg/dL, Potassium, Sodium, Alanine transaminase, Aspartate transaminase, and a class label indicating the presence/absence of COVID-19 disease. In this study, the dataset consists of 520 data samples that belongs to normal or healthy persons labelled as value 0 and 80 COVID-19 infected patients data samples labelled as value 1. The dataset is highly imbalanced in the ratio (6.5:1) with 86.67% of data samples belonging to healthy patient records and 13.33% data samples belonging to the infected category.

3.2. DATA PREPROCESSING

Data preprocessing techniques enhance generalisation and reduce overfitting, which improves the performance of classification algorithms. Data standardisation is a preprocessing method that prepares the dataset prior to its use in classification tasks. Common data normalization techniques include min-max normalization and zero-mean (Z-score) normalization. In this study, the Z-score normalization technique was applied to the entire dataset, resulting in a zero mean and a standard deviation of 1. The data were subsequently scaled to the range of (-1, +1), as shown in Equation 1.

$$Z_{score} = \frac{(x - \bar{x})}{\sqrt{\frac{\sum(x - \bar{x})^2}{N}}} \quad (1)$$

where, x represents the value of the feature in a data sample, $\bar{x}$ is the average value, the denominator represents the standard deviation of a feature and N denotes the number of samples in the dataset.

Table 1: Clinical and Laboratory findings used for diagnosis of COVID-19 infection

Feature Name	Range	Type
Haematocrit (HCT)	-4.50 - 2.66	Numeric
Red blood Cells (RBC)	-3.97 - 3.64	Numeric
Basophils	-1.14 - 11.08	Numeric
Monocytes	-2.16 - 4.53	Numeric
Eosinophils	-0.83 - 8.35	Numeric
Urea	-1.63 - 11.24	Numeric
Protein C-reativa mg/dL (CPR)	-0.53 - 8.02	Numeric

Creatinine(CR)	-2.38 - 5.05	Numeric
Haemoglobin (Hgb)	-4.34 - 2.67	Numeric
Potassium (K)	-2.28 - 4.10	Numeric
Sodium (Na)	-5.24 - 4.09	Numeric
Leukocytes (WBC)	-2.02 - 4.52	Numeric
Lymphocytes	-1.86 - 3.76	Numeric
Neutrophils	-3.33 - 2.53	Numeric
Platelet(PLT)	-2.55 - 9.53	Numeric
Serum Glucose	-1.10 - 7.01	Numeric
Alanine transaminase (ALT)	-0.64 - 7.93	Numeric
Aspartate transaminase (AST)	-0.70 - 7.23	Numeric
Label	Yes	No Categorical

3.3. FEATURE SELECTION ALGORITHM

Feature selection techniques are utilized to select significant and relevant feature attributes that helps to improve the performance of the machine learning algorithm. In this study, we have utilized mutual information (MI) technique [23]. It is a statistical measure for selecting significant feature attribute by measuring the correlation between the variables. The computed MI score is greater than or equal to zero. The higher the MI score then higher the correlation between two variables, when MI score is equal to zero, then there is no correlation and the variables are independent. The MI measuring the statistical dependence between two random variables X and Y with N observations is given in equation.

$$I(X; Y) = H(Y) - H(Y|X) \quad (2)$$

where H(Y) represents the entropy of Y that quantifies the expected uncertainty level in a discrete random variable Y.

$$H(Y) = -\sum_{y \in Y} P(y) \log(P(y)) \quad (3)$$

The joint entropy between two discrete random variables X and Y is computed as given in equation 4.

$$H(X, Y) = -\sum_{x \in X} \sum_{y \in Y} P(x, y) \log(P(x, y)) \quad (4)$$

where P(x,y) is the joint probability mass function of random variable X and Y.

The conditional entropy H(Y|X) is defined as the entropy of random variable Y by observing the random variable X which is calculated using equation 5.

$$H(X|Y) = -\sum_{x \in X} \sum_{y \in Y} P(y|x) \log(P(y|x)) \quad (5)$$

P(x); P(y) represents the marginal probability mass functions of random variable X and Y, P(y|x) is the conditional probability of variable Y given variable X. The mutual information between two variables X and Y, a positive value signifies there is a dependency between the variables and a value of zero indicates the variables are independent. In this study, X signifies the number of independent feature attributes and Y is a dependent value that signifies the target class variable. From the figure 1, it can be seen that only eleven features attributes have value greater than zero. Hence, we have selected top 9 variables namely Leukocytes, Platelets, Eosinophils, Creatinine, Hemo globin, Basophils, Monocytes, Red blood Cells, Proteina C reativa mg/dL having high MI scores for classification of COVID-19 disease.

3.4. COST-SENSITIVE CLASSIFICATION

In medical diagnosis, most of the real-world datasets used for analysis are highly imbalanced. A cost-sensitive learning is a technique that considers the misclassification cost to minimize misclassification rate. The misclassification cost plays a important part in cost-sensitive learning to handle the imbalanced distribution of classes. In this study, we have utilized cost-sensitive meta learning technique that integrates cost-sensitiveness in the cost-insensitive classifier algorithms. In binary classification problem, given a sample xi such that $x_i \in X$ having N samples consisting of k feature attributes with two classes $c_i \in \{0, 1\}$.

The main objective is to identify or predict the target class $y_i \in \{0, 1\}$ of a sample. The principle of minimum expected cost states that the objective of a cost sensitive classification is to classify a sample into a class having minimum expected cost. The expected cost or risk $R(i|x)$ associated with binary classification in predicting a sample x into a target class i can be expressed as in equation 6.

$$\arg \min R(i|x) = \arg \min \sum_{j \neq i} \text{Prob}(j|x) \text{Cost}_{(j,i)} \quad (6)$$

where $R(i|x)$ is the expected risk of classifying sample x into class i , $\text{Prob}(j|x)$ is the posterior probability of sample x in class j , $\text{Cost}_{(j,i)}$ is the misclassification cost or risk associated with a sample belonging to class j misclassified as class i .

In cost-sensitive binary classification, the samples of minority class will be proportionately given a higher weight than the majority class. Without loss of generality, here, it is assumed that the negative class (N) is considered as the majority class whereas the positive class (P) as the minority class. In an imbalanced classification, a 2×2 cost matrix is constructed in cost-sensitive learning such that the cost of $\text{Cost}_{(0,1)} \neq \text{Cost}_{(1,0)}$ and $\text{Cost}_{(0,0)} = \text{Cost}_{(1,1)} = 0$. A weighting-based cost-sensitive learning which is a sampling method that allocates a normalized weight for each samples is used in this study. Let $W_{\min}$ be the cost weight assigned to the instances belonging to minority class and W_{maj} be the cost weight given to the majority class, then the misclassification cost is a normalized weight assigned as given in equation 7. The weight matrix is constructed as given in table 2.

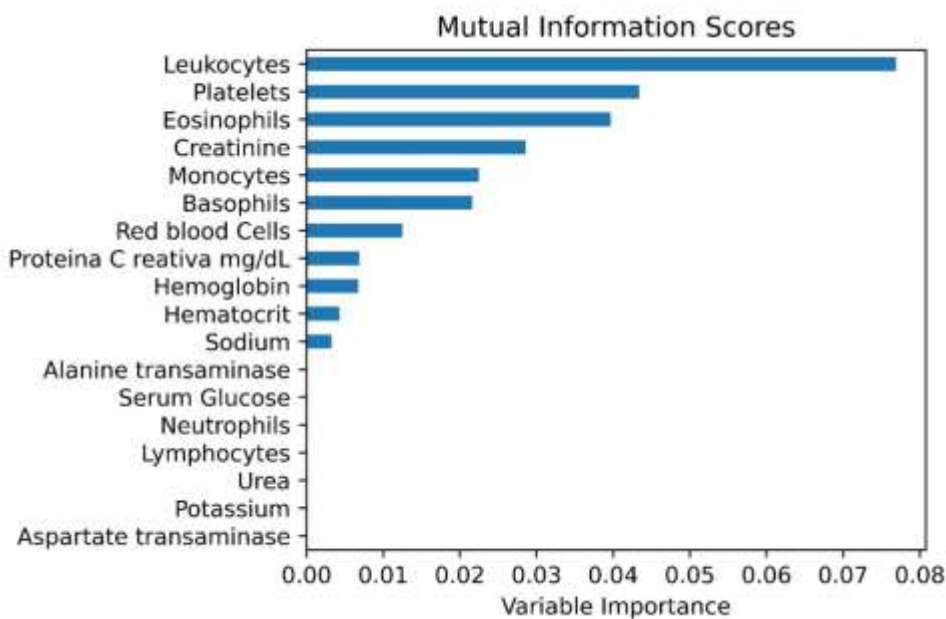


Figure 1: Feature Selection based on Mutual Information

$$\text{Cost}_{(1,0)} = \text{FN} = W_{\min}$$

$$\text{Cost}_{(0,1)} = \text{FP} = W_{\text{maj}} \quad (7)$$

where FN and FP are the false negative rate and false positive rate respectively, $W_{\min}$ and W_{maj} are the weights of minority and majority class respectively and $\text{Cost}_{(1,0)}$ and $\text{Cost}_{(0,1)}$ are the misclassification cost of positive and negative classes respectively. The weight of minority class $W_{\min}$ is greater than the weight of the majority class W_{maj} (ie., the cost incurred for misclassification of minority class is higher than the cost of majority class, $\text{Cost}_{(1,0)} > \text{Cost}_{(0,1)}$) for imbalanced classification.

Table 2: Weight Matrix used in the cost-sensitive classification of COVID-19

	Predicted Positive	Predicted Negative
Actual Positive	0	$\text{Cost}_{(1,0)}$ or $W_{\min}$
Actual Negative	$\text{Cost}_{(0,1)}$ or W_{maj}	0

3.5. PROPOSED COST-SENSITIVE HETEROGENEOUS STACKED ENSEMBLE CLASSIFIER MODEL (CS- HSEC) FOR PREDICTING SARS-COV-2 INFECTION

The proposed cost-sensitive, stacked ensemble classifier model is applied for a binary classification to predict SARS-CoV-2 infection using routine laboratory findings. Stacked ensemble classifier model is an ensemble model that combines several similar or different base classifiers to construct a multi-level stacking architecture with a meta-classifier. The proposed stacked ensemble classifier is a heterogeneous classifier model that combines prediction of different CS-base classifiers. In this study, we have constructed a stacked ensemble model using three different cost-sensitive base classifiers (CS-base classifiers) namely RF, GBM and KNN with Naive Bayes classifier as meta-learner. A weighting-based cost-sensitive learning is incorporated that applies a normalized weight to each class based on their distribution. The predictive results of the cost-sensitive base classifiers are used as the input attributes to train the meta-classifier. The proposed cost-sensitive heterogeneous stacking classifier (CS-HSEC) applies stratified 10-fold cross-validation with data shuffling to improve the classification accuracy and to avoid overfitting.

Let us consider the dataset $D=\{X,Y\}$ having N sample of individuals with k independent feature attributes and each sample $x_i \in X$ say $x_i = \{x_1, x_2, \dots, x_k\}$ belonging to one of m classes such that $y_i \in Y = \{y_1, y_2, \dots, y_{km}\}$. The dataset D is partitioned into two namely training D_{Tr} and testing D_{Te} . During the training process, the CS-base classifiers are trained using the D_{Tr} and the stacked ensemble classifier combines the prediction results of the CS-base classifiers. A K -fold cross-validation is applied to predict the COVID-19 infection using the Naive Bayes as meta classifier by dividing the D_{Tr} into K disjoint subsets that is stratified to maintain same class distribution for each fold during model training and validation.

During the training process, each CS-base classifiers are trained using $K-1$ subsets and one subset is used for validation. The process is repeated K times for all the three base classifiers and all the predicted results of validation subset of three different base classifiers (three prediction results) are combined along with the actual target class to form the meta-data of dimension $(N \times 4)$ as a new training set. The new training dataset or meta-data is used as the training data to train the Naive Bayes meta classifier. The prediction result of the different CS-base classifiers can be a binary class label or probability values. We have taken the probability of prediction as result from three CS-base classifiers for validation. Finally, in the testing phase, we make prediction on the test dataset D_{Te} to evaluate the performance of the CS-base classifiers. The predictive result (probabilities) of each base classifiers are combined with the actual class and used by the meta-classifier for model evaluation. The prediction result of the meta classifier is the final prediction result of the cost-sensitive stacked ensemble algorithm.

4. PERFORMANCE EVALUATION AND RESULTS

4.1. EVALUATION MEASURES

The performance of the classification algorithm is assessed primarily in terms of accuracy. For imbalanced datasets, additional metrics are considered, including recall, specificity, precision, F1-score, Cohen's kappa coefficient, Matthews correlation coefficient, and Area Under the Curve, since the majority class can significantly influence overall accuracy. Metrics derived from the confusion matrix, such as True Positive rate (TP), False Positive rate (FP), True Negative rate (TN), and False Negative rate (FN), are also utilized to evaluate classifier performance. The accuracy of the classifier is calculated using equation (8).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (8)$$

Recall, or TPR, is an important metric in imbalanced classification of a medical dataset because it evaluates the classifier's performance in identifying the minority class. Specificity is used to evaluate a classifier's effectiveness in identifying majority-class samples. The sensitivity and specificity are measured using equation (9) and equation (10), respectively

$$Recall = \frac{TP}{TP+FN} \quad (9)$$

$$Specificity = \frac{TN}{TN+FP} \quad (10)$$

Similarly, precision, F1 score and the quality of a binary classifier measured with Matthew's Correlation coefficient (MCC) are calculated using equation(11) - (13) respectively.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (11)$$

$$F1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (12)$$

$$MCC = \frac{TP+TN-FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (13)$$

The kappa statistical test (KS) that measures the inter-rater reliability of classification accuracy of a classifier considering the random accuracy of the classifier is given in equation (14).

$$k = \frac{2 \times (TP \times TN - FP \times FN)}{(TP+FP) \times (FP+TN) + (TP+FN) \times (FN+TN)} \quad (14)$$

4.2. EXPERIMENTAL RESULTS

The classifiers were evaluated on an 11th Generation Intel Core i5-1135G7 2.40GHz Windows computer with 16 GB of main memory. Classifier models were implemented in Python using Google Colaboratory. The pandas library facilitated dataset loading [30], scikit-learn [31] supported machine learning algorithms, and matplotlib enabled data visualisation [32]. The performance of the proposed cost-sensitive heterogeneous stacked ensemble model, as well as other classifier models, was evaluated to assess their effectiveness in predicting COVID-19 infection using clinical tests and laboratory findings.

The proposed algorithm was evaluated using stratified 10-fold cross-validation with the cross_val_score function from scikit-learn for COVID-19 detection. The CS-HSEC method achieved a classification accuracy of 90.01%, a sensitivity of 0.75, and a specificity of 0.92 in predicting SARS-CoV-2 infection. Figure 2 displays the classification accuracy of the CS-HSEC and other developed classifier models, assessed by stratified K-fold cross-validation, in the form of a box plot.

Figure 3 presents the receiver operating characteristic curve of the proposed CS-HSEC model for stratified 10-fold cross-validation. Although 10-fold cross-validation is widely used in medical diagnosis for relatively small datasets, it may result in less clear interpretation of results in medical applications. Therefore, a train-test split model was also applied to evaluate the classification performance of the classifiers in predicting COVID-19 infection.

The dataset was split using a stratified, shuffled train-test split with the train-test-split command from scikit-learn. The resulting data included 90% training data with 520 samples, comprising 468 normal samples and 72 COVID-19 infected samples. The test dataset consisted of 60 samples, with 52 normal and 8 abnormal samples. Table 3 presents the average performance results over 20 iterations on different datasets for each classifier. According to Table 3, the proposed cost-sensitive heterogeneous stacked ensemble classifier achieved 33% sensitivity and 93% accuracy, 13% sensitivity, 99.2% accuracy, and 3% specificity, with an Area Under the Curve (AUC) of 0.96. Additionally, the cost-sensitive stacked ensemble classifier attained an F1 score of 0.97, indicating low false positive and negative rates. The proposed model also demonstrated high precision of 0.96 and achieved a Kappa statistical value and Matthews Correlation Coefficient of 0.93, respectively.

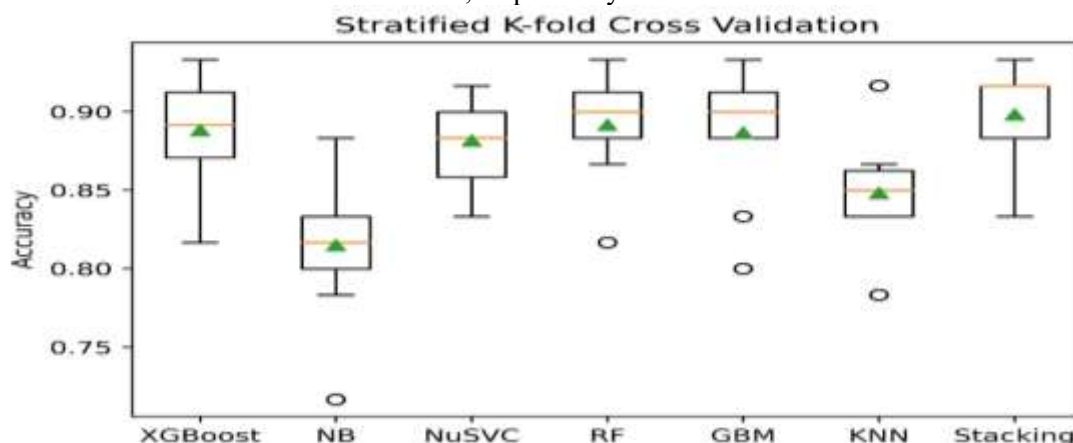


Figure 2: Box-plot showing the stratified K-fold cross-validation accuracy of the classifiers

Table 3: Classification results of 20 simulation runs obtained using stratified train and test split for the prediction of COVID-19 infection from laboratory findings.

Algorithm	CA%	Recall	Spec	Prec.	F1	Kappa	MCC	AUC
CS-XGB	93.42	0.54	0.99	0.95	0.68	0.65	0.68	0.77
NB	87.33	0.73	0.89	0.53	0.61	0.54	0.55	0.81
CS-NuSVC	92.00	0.55	0.98	0.80	0.64	0.59	0.61	0.76
CS-KNN	93.50	0.62	0.98	0.89	0.72	0.69	0.70	0.80
CS-RF	94.75	0.68	0.99	0.91	0.78	0.75	0.76	0.85
CS-GBM	96.67	0.80	0.99	0.95	0.86	0.84	0.85	0.90
CS-Stacked Ensemble	98.33	0.93	0.99	0.96	0.94	0.93	0.93	0.96

CA - Accuracy, Spec - Specificity, Prec - Precision, MCC - Matthew's Correlation Coefficient

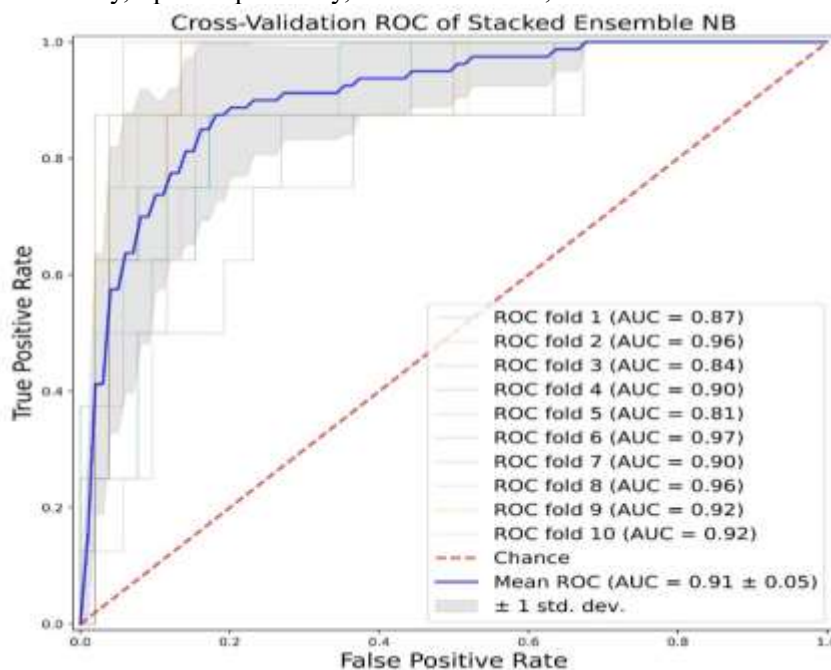


Figure 3: ROC curve of proposed cost-sensitive heterogeneous stacked ensemble classifier with stratified 10-fold cross validation

Figure 4 presents the receiver operating characteristic (ROC) curves for all classifiers evaluated on the COVID-19 dataset. The area under the curve (AUC) for the cost-sensitive stacked ensemble classifier demonstrates effective identification of both positive and negative samples. Experimental results in Table 3 highlight the robustness of the proposed classifier in comparison with other developed models. Table 4 provides a summary of previous studies and the machine learning algorithms employed, alongside the proposed CS-HSEC model, for COVID-19 prediction using laboratory test results from Hospital Israelita Albert Einstein in Sao Paulo, Brazil. As indicated in Table 1, previous studies have not utilized a stacked ensemble algorithm for COVID-19 prediction, except in [] and [25]. To the best of the authors' knowledge, this model is the first to apply cost-sensitive learning for COVID-19 prediction using clinical and laboratory data. The CS-HSEC algorithm outperforms all other developed classifier models, with the exception of [16], which exhibits a slight reduction in accuracy. Only nine feature attributes were utilized, thereby reducing computational cost compared to other classifiers, while achieving 98.33% accuracy in COVID-19 prediction.

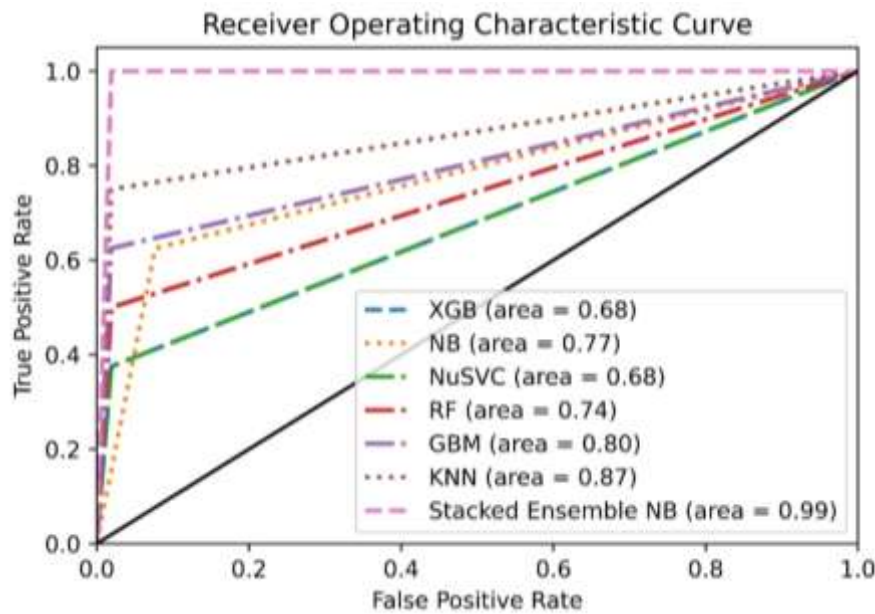


Figure 4: ROC curve of the classifier models for a single run using train-test split.

Table 4: Comparing the results of proposed algorithm with existing studies in literature for predicting COVID-19 infection using laboratory test results.

Method	#Features	Acc(%)	Sen(%)	Spec(%)	AUC(%)
ANN, CNN, RNN, LSTM, CNNLSTM, CNNRNN [15]	18	92.30	93.68	NR	90
RF,LR, GLMNET, ANN [17]	14	87	65	91	86
NN, RF, XGBoost, LR, SVM [20]	15	NR	68	85	84.7
XMLP, SVM, RT, RF, BN, NB + SMOTE [18]	24	95.16	96.8	93.60	NR
Ensemble SVM Models + SMOTE [25]	16	NR	70.25	85.98	86.78
LR, NN, RF, SVM, XGB [27]	97	95.1	80	98.6	96
ANN, CNN, RNN + SMOTE [29]	18	94.95	94.98	NR	100
Ensemble RF, ET, LR with XGBoost with SMOTE [16]	18	99.88	98.72	99.9	98.72
Proposed CS-HSEC	9	98.33	93.13	99.23	96.1

#Features-Number of Features selected, NR- Not Reported The limitations of the proposed cost-sensitive heterogeneous stacked ensemble classifier model is as follows:

- (1) The result obtained by the proposed CS-HSEC is tested with only 600 samples from a single dataset.
- (2) The performance of the model needs to be tested on other COVID-19 datasets from different regions to improve the generalization ability.

5.CONCLUSION:

Accurate diagnosis and prediction of SARS-CoV-2 infection are essential for timely treatment and disease containment. This study introduces a cost-sensitive heterogeneous stacked ensemble classifier (CS-HSEC) for detecting COVID-19 patients using routine clinical tests and laboratory results from Hospital Israelita Albert Einstein, Brazil. Highly significant and discriminant features were selected

through a hybrid approach incorporating the mutual information (MI) technique to enhance COVID-19 prediction. The heterogeneous stacking ensemble is a two-level model that integrates cost-sensitive learning and leverages structural diversity by employing multiple classifier models. Probabilistic outputs from level-1 classifiers, including random forest, gradient boosting machine, and K-nearest neighbour, are combined with the target class and provided as input to a Naive Bayes meta-classifier at level-2, which substantially improves predictive performance in COVID-19 diagnosis. Model performance was validated using a stratified train-test split and 10-fold cross-validation. Experimental results demonstrate that the CS-HSEC achieves 98.33% accuracy, 93.13% sensitivity, and an AUC of 96.1% with 90% training and 10% testing data. The algorithm requires only ten routine clinical blood tests and laboratory findings to predict COVID-19 patients.

Further research is required to address the challenges associated with implementing artificial intelligence-based clinical decision support systems (CDSS) for the automated diagnosis of COVID-19 infection. Such efforts necessitate rigorous validation in collaboration with clinical and medical experts across diverse ethnic groups and countries. Additionally, future work may involve developing machine learning and deep learning algorithms utilizing other modalities, including chest X-ray images, computed tomography, and lung ultrasound findings, to enhance COVID-19 detection.

References:

- [1]. WHO Director-Generals opening remarks at the media briefing on COVID-19 - 11 March 2020. <https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020>.
- [2]. D.Lin, L.Liu, M.Zhang, Y.Hu, Q.Yang, J.Guo, et al. Evaluations of serological test in the diagnosis of 2019 novel coronavirus (SARSCoV2) infections during the COVID-19 outbreak. *Eur J Clin Microbiol Infect Dis.*,pp.1–7,2020.
- [3]. S.R Nayak, D.R.Nayak, U.Sinha, V. Arora, R.B.Pachori, Application of deep learning techniques for detection of COVID-19 cases using chest X-ray images: A comprehensive study, *Biomedical Signal Processing and Control*,vol.64, pp. 1-12, 2021.
- [4]. D.Li, D.Wang, J.Dong, N.Wang, H.Huang, X.Hu, C.Xia, False-Negative Results of Real-Time Reverse-Transcriptase Polymerase Chain Reaction for Severe Acute Respiratory Syndrome Coronavirus 2: Role of Deep-Learning-Based CT Diagnosis and Insights from Two Cases.*Korean J Radiol*, vol.21, no.4, pp. 505-508, 2020.
- [5]. K.El Asnaoui , Y.Chawki, Using X-ray images and deep learning for automated detection of coronavirus disease. *J Biomol Struct Dyn*, vol.39, no.10, pp.3615-3626, 2021.
- [6]. L.Breiman, Random forests, *Machine Learning* 45, (2001), pp.5–32.
- [7]. M.Imani, Automatic diagnosis of coronavirus (COVID-19) using shape and texture characteristics extracted from X-Ray and CT-Scan images, *Biomedical Signal Processing and Control*, vol.68, pp. 1-12, 2021.
- [8]. Velusamy,D., and Ramasamy,K., Ensemble of heterogeneous classifiers for diagnosis and prediction of coronary artery disease with reduced feature subset,*Computer Methods and Programs in Biomedicine*, 198,(2021), pp. 1-13.
- [9]. J.M.Johnson, T.M.Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, vol.6, no.1, 2019. <https://doi.org/10.1186/s40537-019-0192-5>.
- [10].S.Balasubramanian, R.Kashyap, S.T.Cvn, and M.Anuradha, Hybrid prediction model for type-2 diabetes with class imbalance, In: 2020 IEEE international conference on machine learning and applied network technologies (ICMLANT), Dec. 2020. pp. 1–6. <https://doi.org/10.1109/ICMLANT50963.2020.9355975>.
- [11].I.D.Mienye, Y.Sun,Performance analysis of cost-sensitive learning methods with application to imbalanced medical data, *Informatics in Medicine Unlocked*, vol.25, pp.1-10, 2021
- [12].S.I.Ali , et al. Ensemble feature ranking for cost-based non-overlapping groups: a case study of chronic kidney disease diagnosis in developing countries. *IEEE Access*, vol.8, pp. 215623 – 215648, 2020. <https://doi.org/10.1109/ACCESS.2020.3040650>.
- [13].D.Gan, J.Shen, B.An, M.Xu, N.Liu, Integrating TANBN with cost sensitive classification algorithm for imbalanced data in medical diagnosis. *Comput. Ind Eng*, vol.140, 2020. <https://doi.org/10.1016/j.cie.2019.106266>.

- [14].M.Phankokkrud. Cost-Sensitive extreme gradient boosting for imbalanced classification of breast cancer diagnosis, in 2020 10th IEEE international Conference on control system. Computing and Engineering (ICCSCE) Aug. pp.46–51, 2020. <https://doi.org/10.1109/ICCSCE50387.2020.9204948>.
- [15].T.B.Alakus, and I.Turkoglu, Comparison of deep learning approaches to predict COVID-19 infection, *Chaos Solitons Fractals* (140) (2020) 110120.
- [16].M.AlJame, I.Ahmad, A.Imtiaz, and A.Mohammed, Ensemble learning model for diagnosing COVID-19 from routine blood tests, *Informatics in Medicine Unlocked*, vol. 21, 2020, 100449, <https://doi.org/10.1016/j.imu.2020.100449>.
- [17].A.Banerjee, S.Ray, B.Vorselaars, J.Kitson, M.Mamalakis, S.Weeks, M.Baker, and L.S.Mackenzie, Use of machine learning and artificial intelligence to predict SARS-CoV-2 infection from full blood counts in a population, *Int. Immunopharmacol.* (86) (2020) 106705.
- [18].V.A.de.F.Barbosa, J.C.Gomes, M.A.de.Santana, J.E.de.A.Albuquerque, R.G.de.Souza, R.E.de Souza, W.P.dos Santos W, Heg.IA: an intelligent system to support diagnosis of covid-19 based on blood tests. medRxiv URL, <https://doi.org/10.1101/2020.05.14.20102533>; 2020.
- [19].D.Brinati, A.Campagner, D.Ferrari, M.Locatelli, G.Banfi, and F.Cabitzza, Detection of COVID-19 infection from routine blood exams with machine learning: A feasibility study, *J. Med. Syst.* 44 (8) (2020) 135.
- [20].A.F.de.M.Batista, J.L.Miraglia, T.H.R.Donato, and A.D.P.Chiavegatto, COVID-19 diagnosis prediction in emergency care patients: a machine learning approach, 2020.
- [21].Data4u, E. Hospital Israelita Albert Einstein, Sao Paulo, Brazil, Diagnosis of Covid-19 and its clinical spectrum, 3,(2020) <https://www.kaggle.com/einsteindata4u/covid19>.
- [22].J.Han , and M.Kamber , *Data Mining: Concepts and Techniques* Morgan Kaufmann, 2nd Edition, 2006.
- [23].P.Ganesh Kumar, T.Aruldoss Albert Victoire, P.Renukadevi and D.Devaraj,Design of fuzzy expert system for microarray data classification using a novel Genetic Swarm Algorithm, *Expert Systems with Applications*, vol.39, no.2, pp.1811–1821,2012.
- [24].X.Jiang, M.Coffee, A.Bari, J.Wang, X.Jiang, J.J.Huang, et al., Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity, *Comput. Mater. Continua*, vol.63, no.1, pp.537–551,2020.
- [25].F.Souares, A.Villavicencio, F.S.Fogliatto, M.H.P.Rigatto, M.J.Anzanello, M.Idiart, and M.Stevenson, A novel specific artificial intelligence-based method to identify covid-19 cases using simple blood exams. medRxiv URL, <https://doi.org/10.1101/2020.04.10.20061036>; 2020.
- [26].P.Schwab, A.D.Schütte, B.Dietz, and S.Bauer, Clinical predictive models for COVID-19: Systematic study, *J. Med. Internet Res.* vol.22, no.10, 2020. e21439.
- [27].X. Mei, H.C. Lee, K. Diao, M. Huang, B. Lin, M.C. Liu, et al., Artificial intelligence-enabled rapid diagnosis of patients with COVID-19, *Nat. Med.*, vol.26, pp.1224-1228,2020. <http://dx.doi.org/10.1038/s41591-020-0931-3>.
- [28].V.Goreke, V.Sari, and S.Kockanat, A novel classifier architecture based on deep neural network for COVID-19 detection using laboratory findings, *Appl. Soft Comp.*, vol.106, 2021. 107329.
- [29].W.McKinney, et al., *Pandas: a foundational python library for data analysis and statistics*. Python for High Performance and Scientific Computing, vol.14, no.9, 2011.
- [30].F.Pedregosa, G.Varoquaux, A.Gramfort, V.Michel, B.Thirion, O.Grisel, M.Blondel, P.Prettenhofer, R.Weiss, V.Dubourg, et al., *Scikit-learn: Machine learning in python*. The Journal of Machine Learning research, 12:2825–2830, 2011.
- [31].J.D.Hunter, *Matplotlib: a 2d graphics environment*. Computing in science and engineering, vol.9, no.3, pp.90–95, 2007.