JIRSS (2025)
Vol. 24, No. 2, pp 16-32
DOI: 10.22034/jirss.2025.2053767.1102

Statistical Learning in the Fight against COVID-19: A Focus on Diagnosis

Mostafa Tamandi[1], Zahra Kamiab[2], Fatemeh Bahremand[3,4], Mohammad Yasin Zamanian[5,6], Mohammadreza Gholamrezapour[3,4], Seyed Mohammad Ebrahim Pourhosseini[3,4], Gholamreza Bazmandegan[*7,8]

[1]Department of Statistics, Vali-e-Asr University of Rafsanjan, Rafsanjan, Iran.
[2]Department of Community Medicine, School of Medicine, Rafsanjan University of Medical Sciences, Rafsanjan, Iran.
[3]Clinical Research Development Unit, Ali-Ibn Abi-Talib Hospital, Rafsanjan University of Medical Sciences, Rafsanjan, Iran.
[4]Department of Internal Medicine, Ali-Ibn Abi-Talib Hospital, School of Medicine, Rafsanjan University of Medical Sciences, Rafsanjan, Iran.
[5]Department of Physiology, School of Medicine, Hamadan University of Medical Sciences, Hamadan, Iran.
[6]Department of Pharmacology and Toxicology, School of Pharmacy, Hamadan University of Medical Sciences, Hamadan, Iran.
[7]Physiology-Pharmacology Research Center, Research Institute of Basic Medical Sciences, Rafsanjan University of Medical Sciences, Rafsanjan, Iran.
[8]Department of Physiology and Pharmacology, School of Medicine, Rafsanjan University of Medical Sciences, Rafsanjan, Iran.

Received: 18/02/2025, Accepted: 15/12/2025, Published online: 31/01/2026

Abstract. The accurate diagnosis of infectious diseases such as COVID-19 requires statistically reliable classification methods capable of handling complex, heterogeneous, and imbalanced data. In this study, several statistical and machine learning algorithms –Logistic Regression, Linear Discriminant Analysis, K-Nearest Neighbors, Decision

M. Tamandi (tamandi.m@gmail.com).
Z. Kamyab (dr.kamiab89@gmail.com).
F. Bahremand (dr.bahrehmand@yahoo.com).
M. Y. Zamanian (mzamaniyan66@yahoo.com).
M. Gholamrezapour (drmgh49@yahoo.com).
S. M. E. Pourhosseini (s.mepoorhoseini@gmail.com).
[*]**Corresponding Author:** G. Bazmandegan (bkhrbster@gmail.com).

Tree, and Random Forest –were comparatively evaluated using clinical and laboratory data from 506 hospitalized patients in Rafsanjan, Iran. The dataset included 27 categorical and 11 quantitative variables. To address class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was employed. Model performance was assessed using a comprehensive set of criteria, including accuracy, sensitivity, specificity, positive and negative predictive values (NPV), and the area under the ROC curve. The comparative analysis showed that Random Forest and Logistic Regression achieved the best overall performance, while SMOTE improved sensitivity and NPV at the expense of specificity. The findings emphasize the importance of appropriate imbalance correction and multi-metric evaluation in developing statistically robust diagnostic models for medical data.

Keywords. COVID-19, PCR test, SMOTE, Random Forest, Logistic Regression.

MSC: 62H30, 62P10, 68T09.

1 Introduction

Classification has long played a pivotal role in medical research, forming the foundation for many diagnostic and prognostic systems. It involves assigning observations to predefined categories based on a set of measured attributes, thereby allowing the prediction of qualitative outcomes that characterize each patient. In healthcare diagnostics, this predictive capability is particularly vital, as it supports clinicians in identifying disease patterns and improving decision-making (Organization and Statistics, 1980). Broadly, classification methods can be divided into two categories: parametric and nonparametric models (Malik and Munjal, 2021). Parametric approaches, grounded in classical statistical theory, rely on explicit assumptions about the underlying data distribution. Techniques such as Logistic Regression (LR) and Linear Discriminant Analysis (LDA) have long served as cornerstones in medical research, providing interpretable models that quantify relationships between predictors and health outcomes (Schober and Vetter, 2021). In contrast, nonparametric or machine learning–based models, such as K-Nearest Neighbors (KNN), Decision Tree (DT), and Random Forest (RF), offer more flexible frameworks that do not require distributional assumptions. These algorithms are adept at uncovering complex, nonlinear patterns within heterogeneous medical datasets (Rashidi and Hu, 2023), thereby complementing traditional statistical approaches in modern healthcare analytics.

Despite substantial advances in both statistical and machine learning methodologies, significant challenges remain when these techniques are applied to real-world data. One of the most critical challenges is class imbalance, a situation in which the number of observations in one class markedly exceeds that of another. Such imbalance can bias the training process, causing classifiers to favor the majority class and yielding misleadingly high overall accuracy while failing to correctly identify minority outcomes.

In the context of medical disease diagnosis, where individuals are classified as patients or non-patients, class imbalance is a common and well-recognized challenge, as diagnostic data frequently exhibit disproportionate class distributions. This imbalance typically arises because the number of affected individuals is often substantially smaller than that of unaffected individuals in clinical datasets. Moreover, false-negative predictions in such diagnostic settings may lead to serious clinical and public health consequences, including delayed diagnosis, inappropriate treatment, and increased disease transmission or progression.Therefore, effectively addressing class imbalance is essential to ensure reliable and clinically meaningful classification performance. To address this challenge, the Synthetic Minority Over-sampling Technique (SMOTE) has been widely developed and applied in the literature (Chawla and Kegelmeyer, 2002). SMOTE mitigates class imbalance by generating synthetic samples for the minority class through interpolation between existing observations, thereby improving the representativeness of underrepresented classes during model training. Unlike simple duplication, this approach enhances the representativeness of the minority class, improving model generalization and yielding more reliable estimates of key performance metrics such as sensitivity, specificity, and predictive values.

Building on these theoretical foundations, the present study employs a combination of parametric and nonparametric classification approaches to address the challenges inherent in medical disease diagnosis. Although the empirical analysis is motivated by COVID-19—a condition whose diagnosis has posed substantial challenges to clinicians, particularly in the presence of imbalanced clinical data—the proposed framework is not disease-specific and can be readily extended to other diagnostic contexts. By comparing Logistic Regression (LR), Linear Discriminant Analysis (LDA), k-Nearest Neighbors (KNN), Decision Tree (DT), and Random Forest (RF) models using data from hospitalized patients, this study evaluates their predictive performance, robustness to class imbalance, and clinical interpretability. Through this comparative analysis, the study underscores the complementary strengths of traditional statistical methods and modern machine learning algorithms, highlighting their combined potential to support and enhance diagnostic decision-making across a broad range of healthcare applications.

Ultimately, this investigation underscores the growing importance of statistical learning in medical data analysis and the need for methodological strategies—such as SMOTE—to ensure fairness and reliability in model evaluation. As the global community continues to confront the challenges of infectious diseases like COVID-19, developing accurate, interpretable, and statistically sound diagnostic models remains a cornerstone for improving public health outcomes and advancing evidence-based medicine (Sidey-Gibbons and Sidey-Gibbons, 2019; Fischhoff, 2020). The novelty of our study lies in its integrated evaluation framework, which combines traditional statistical learning and modern machine learning techniques under class imbalance correction to provide a unified, statistically rigorous comparison of diagnostic performance. By incorporating SMOTE and a multi-metric assessment strategy, this work contributes new insights into the reliability, interpretability, and robustness of classification models for real-world medical data. The remainder of this paper is organized as follows.

Section 2 describes the data sources, preprocessing steps, and the methodology used for model development and evaluation. Section 3 presents the empirical results, including performance metrics for each classifier under both original and SMOTE-adjusted data. Section 4 discusses the implications of the findings in comparison with previous studies, and Section 5 concludes the paper with remarks on the methodological contributions and potential directions for future research.

2 Methods

This section provides a summary of each model employed in the classification of the COVID-19 data. We avoid presenting mathematical formulas or technical algorithms of each method. Readers can refer to the sources indicated for each of them for further study.

Logistic Regression (LR) is a parametric classification method used to build forecasting models based on the analysis of input data. This model is a particular case of a multiple regression model with a binomial response variable. It estimates the conditional probability of allocating a person to each of the pre-specified classes. So, an observation is classified into one of two groups if this probability is greater than 0.5. The LR method leads to a linear classifier whose accuracy depends on some strong assumptions. For example, multicollinearity problems can increase the misclassification error rate. Moreover, the LR model has some unknown parameters which should be estimated by maximum likelihood method or so on. When there are several categorical variables in the set of input variables, these estimates may have wide and imprecise confidence intervals (Weisberg, 2005).

Linear Discriminant Analysis (LDA) is another parametric statistical learning method for classification. In this method, the classification algorithm is converged to a classifier by minimizing the cost of misclassification of training data. In LDA, we must have two strong assumptions. The distribution of input variables must be Gaussian, and the classes should be homogeneous. Since in many applications, such assumptions are difficult to achieve, it should be used carefully (Johnson and Wichern, 2002).

K-Nearest Neighbors (KNN) is a nonparametric learning method designed to model classification experiments. In fact, the KNN classifies k nearest similar values in a group, in which these k values are namely nearest neighbors. In this context, to find the nearest values, a similarity measure such as Euclidian distance is adopted. Then, with an iterative algorithm, observations are classified into two or more groups. In contrast to the simplicity of this supervised learning, the computational cost may be a disadvantage (James and Tibshirani, 2013).

Decision Tree (DT) is a supervised learning model which involves stratifying or segmenting the predictor space into several simple regions. In the DT, we predict that each observation belongs to the most commonly occurring class of training observations in the region to which it belongs. So, a tree is designed with some leaves and nodes. Each observation can be classified into a terminal node of the DT with a conditional

probability given by input variables. To grow a classification tree and stop the DT algorithm, misclassification error rate or Gini index can be used as criteria. Trees can be displayed graphically and so are very easy to explain to non-expert people. But, one of the disadvantages of DT, which is noted by some authors, is that trees generally do not have the same level of predictive accuracy as the parametric classification approaches (James and Tibshirani, 2013).

Random Forest (RF) is achieved by aggregating many decision trees. This machine learning algorithm can solve the problem of the predictive accuracy of the DTs. The RF algorithm works as follows: Several decision trees are built by many bootstrapped training samples. In each iterative, a random sample of m<p input variables is chosen as split candidates from the full set of p inputs. Finally, the algorithm is stopped when some accuracy criterion is achieved. Since this forest grows by uncorrelated trees chosen via bootstrap samples, the accuracy of this classification method is naturally more than DT (Wager and Athey, 2018).

The LDA and LR produce linear decision boundaries whereas no assumptions are made about the shape of the decision boundary in the other studied models. Therefore, we can expect nonparametric approaches to dominate LDA and LR when the decision boundary is highly non-linear. On the other hand, KNN does not tell us which predictors are important while both DT and RF are constructed via the fact that how much each variable can increase the precision of classification. Therefore, each of these methods, parametric or nonparametric, can help us on investigating the nature of classes and using a suitable learning.

2.1 Dataset and Ethics

The study population comprised consecutive patients admitted to Ali-Ibn Abi-Talib Hospital, Rafsanjan between April 26, 2021 and April 30, 2022 with suspected COVID-19 and available admission clinical/laboratory data. Inclusion criteria were: (1) admission for suspected COVID-19 during the study period, and (2) availability of basic admission vitals and laboratory results. Exclusion criteria were duplicate records and records with missing outcome (PCR test) or key predictors. The institutional review board of Rafsanjan University of Medical Sciences approved this retrospective study (Approval No. IR.RUMS.REC.1400.001). Patient consent was waived due to the retrospective and anonymized nature of the data.

2.2 Data preprocessing

The original dataset contained clinical, demographic and basic laboratory variables collected at hospital admission. We inspected missingness and variable distributions before modeling. For numerical variables with low missingness (< 20%), missing values were imputed by the median; for variables with higher missingness we performed sensitivity checks using multiple imputation (mice) — results were not materially different and therefore median imputation is reported in the main analysis. Categorical predictors were kept as factors when modelling with tree-based methods (e.g., RF), and

converted to dummy (one-hot) variables when required (e.g., Logistic regression). Continuous predictors were centered and scaled prior to distance-based models (KNN). Outliers were identified using the interquartile range (IQR) rule. As the number of extreme observations beyond $3 \times IQR$ was very small, we conducted a sensitivity analysis in which these points were excluded. Results remained consistent, suggesting that the presence of outliers did not materially affect model estimates. Moreover, variables potentially observed after PCR result or not available at admission were excluded from predictive modeling to avoid information leakage.

2.3 Model training and validation

To ensure robust evaluation, we split the dataset using a stratified 75/25% train - test partition, preserving the class ratio. All model selection and hyperparameter tuning were carried out using stratified repeated 10-fold cross-validation (10 folds $\times$ 3 repeats) on the training set. Hyperparameters were selected by grid search: KNN ($k \in \{3 \ldots 21\}$), Random Forest (mtry and ntree grid), Decision Tree (cost-complexity pruning using cp from rpart's cptable). For feature selection in regression we used LASSO with 10-fold CV to select predictors; the final LR model reported uses LASSO-selected predictors. Table 1 shows a summary of these methods and their results.

Table 1: Hyperparameter tuning ranges and selected values for each model

Model	**Hyperparameters explored (grid search)**	**Selection method**	**Final selected value**
LR (LASSO)	λ chosen from $\lambda_{\min}$ by 10-fold CV in `glmnet`	Minimum mean CV deviance	$\lambda = 0.017$ ($\lambda_{\min}$)
LDA	Assumes equal covariance across groups	–	–
KNN	$k \in \{3, \ldots, 21\}$	Stratified 10-fold CV (3 repeats)	$k = 7$
DT	cp (complexity parameter) chosen from `rpart`	Minimum cross-validated error	$cp = 0.012$
RF	$mtry \in \{1, \ldots, 15\}$; $ntree \in \{100, \ldots, 1000\}$	Grid search with repeated CV	$mtry = 6$, $ntree = 150$

3 Main results

The final dataset has 506 records with 348 positive and 158 negative PCR tests. The input variables encompassed 27 categorical and 11 quantitative attributes. The output variable is the status of patients in two categories: The patients who truly have Covid-19 (Positive PCR test) and the patients that have some symptoms but whose PCR test is Negative. This variable makes two disjoint classes in our classification analysis. The attributes that are measured from these patients with their values are shown in Table 2.

Figure 1 shows some graphical informations on gender and the results of the PCR test. On the left plot, we have the distribution of age in two female and male groups. As we can see, the number of females with ages above 60 is more than males. The right panel shows the histogram of test results on two gender groups. The percentage of women with positive PCR tests is more than males.

Figure 2 shows the box plots for WBC, LYM, NUT, and O2sat variables in two groups of test results. The number of WBC in patients with negative tests is clearly more than in those with positive tests. A similar result is achieved for the NUT variable.

Table 2: The studied attributes from the COVID-19 patients with their values

Attributes	Values	Attributes	Values
Gender	Female or Male	Systole	57–220
Age	1–101	Diastole	38–116
Nationality	Iranian or other	O2sat	39%–100%
Cough	Yes or No	Cardiovascular	Yes or No
Sputum	Yes or No	Diabetes	Yes or No
Asthma	Yes or No	Pressure	Yes or No
Anorexia	Yes or No	Respiratory disease	Yes or No
Decreased consciousness	Yes or No	Hyperlipidemia	Yes or No
Sore	Yes or No	Kidney	Yes or No
Headache	Yes or No	Dialyze	Yes or No
Fever	Yes or No	Disease	Yes or No
Nausea	Yes or No	Chemotherapy	Yes or No
Vomit	Yes or No	Cancer	Yes or No
Diarrhea	Yes or No	Immunodeficiency	Yes or No
Chills	Yes or No	Hb	4.6–21.6
Myalgia	Yes or No	LYM	0.03–0.93
Weakness	Yes or No	NUT	0.05–0.98
Temperature	35.5–40.6	WBC	1900–128000
Breath	10–72	Pulse	36.5–170

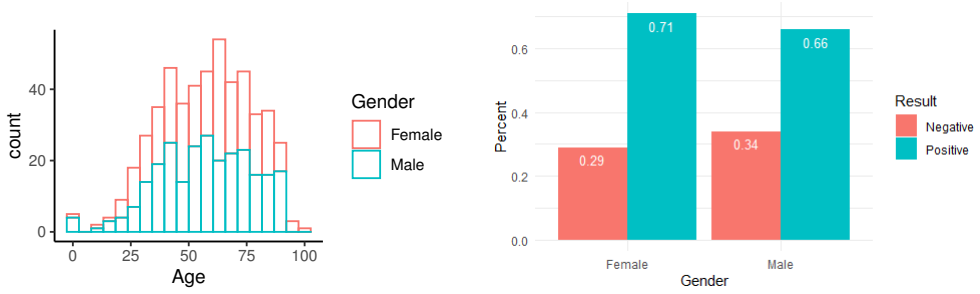

Figure 1: Left: The distribution of age in two gender groups. Right: The histogram of the PCR test's results on two genders.

But the number of LYM in patients with negative tests is less than in those with positive tests. There is not a big difference between the two groups in the O2sat variable. After preproccesing of the data, the model is trained by the training set and the accuracy of this model is tested by the testing set. If such a model shows good precision in predicting output based on the testing set, it can be applied in that experiment. In fact, the model learned by itself in this way. In our study, the training set has 119 negative and 261 PCR positive patient and the test set includes 39 negative versus 86 positive.

In the first experiment, we train the Logistic regression model to make a classification scheme. Table 3 presents the LR coefficients, odds ratios (OR), 95% confidence intervals (CI), and Wald test p-values for the predictors retained after feature selection. Several variables were significantly associated with PCR positivity. Asthma (OR = 1.95, 95% CI: 1.22–3.14, p = 0.005) and fever (OR = 2.05, 95% CI: 1.28–3.27, p = 0.004) increased the odds of testing positive. The presence of anorexia showed a particularly strong effect (OR = 20.6, 95% CI: 4.29–171, p = 0.001), although the wide CI reflects

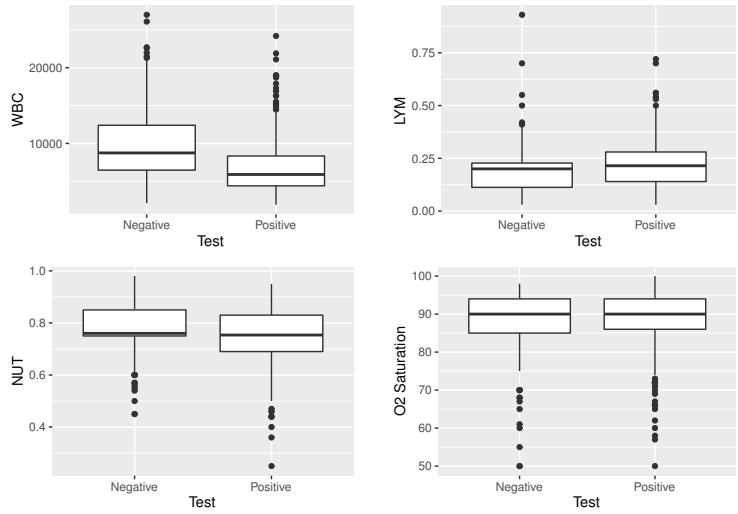

Figure 2: Box plots for WBC, LYM, NUT, and O2Sat variables in COVID-19 patients with positive and negative PCR test results.

Table 3: The estimates of Logistic regression coefficients, odds ratios (OR), confidence intervals for ORs (CI), and p-values

Predictors	**Coef**	**OR**	**CI**	**P-value**
Asthma (Yes/No)	0.666	1.95	(1.22, 3.14)	0.005
Sore (Yes/No)	-2.45	0.085	(0.015, 0.35)	0.0015
Anorexia (Yes/No)	3.02	20.6	(4.29, 171)	0.0010
Fever (Yes/No)	0.71	2.05	(1.28, 3.27)	0.004
Temperature (°C)	-0.61	0.542	(0.395, 0.735)	0.0001
Respiratory (Yes/No)	-1.61	0.198	(0.105, 0.364)	3×10^{-5}
Sputum (Yes/No)	-1.05	0.347	(0.148, 0.781)	0.0120
Kidney (Yes/No)	-0.954	0.385	(0.111, 1.25)	0.115
Consciousness (Yes/No)	-1.32	0.266	(0.117, 0.570)	0.0009
Immunodeficiency (Yes/No)	-1.24	0.287	(0.182, 0.452)	1.02×10^{-6}

limited sample size. In contrast, sore throat, lower temperature, respiratory disease, sputum production, reduced consciousness, and immunodeficiency were associated with significantly reduced odds of PCR positivity (all OR$< 1, p < 0.01$). Kidney disease showed an OR < 1 (0.39) but did not reach statistical significance (p = 0.115). Collectively, these findings indicate that both clinical symptoms (fever, sore throat, anorexia) and comorbidities (respiratory disease, immunodeficiency) play important roles in predicting PCR test outcomes.

For each predictive model, we evaluated performance using a comprehensive set of diagnostic metrics. Accuracy represents the proportion of correctly classified cases among all patients, but can be misleading in imbalanced datasets. Therefore, we additionally report Sensitivity (True Positive Rate), defined as the proportion of PCR-positive patients correctly identified, and Specificity (True Negative Rate), the proportion of PCR-negative patients correctly identified. To reflect clinical utility, we further

report the Positive Predictive Value (PPV), i.e., the probability that a patient predicted positive is truly positive, and the Negative Predictive Value (NPV), the probability that a patient predicted negative is truly negative. Because PPV and NPV depend on class prevalence, they provide complementary information to sensitivity and specificity in applied clinical contexts. Finally, the Area Under the Receiver Operating Characteristic Curve (AUC) was computed to assess overall discriminative ability across all possible classification thresholds, independent of a specific cutoff. This multidimensional evaluation provides a more robust and clinically meaningful assessment of model performance than accuracy or AUC alone.

Since, the dataset was imbalanced, with approximately 69% PCR-positive and 31% PCR-negative patients, we applied the SMOTE during model training. SMOTE is a widely used algorithm that generates synthetic examples of the minority class by interpolating between existing minority instances in feature space (Chawla and Kegelmeyer, 2002). Unlike simple random oversampling, which duplicates minority cases and may lead to overfitting, SMOTE creates new, plausible samples along the line segments joining a minority case and its k nearest minority neighbors. This preserves the distributional structure of the minority class while reducing variance introduced by duplication. In our study, we applied SMOTE only to the training set (never to the test set) to avoid information leakage. All models were trained on the SMOTE-balanced training data and evaluated on the original (unbalanced) test data to provide unbiased estimates of real-world performance.

Table 4 presents the confusion matrix for the Logistic regression model, after SMOTE, detailing the number of true negatives (TN), true positives (TP), false negatives (FN), and false positives (FP). This representation allows readers to directly observe the distribution of correct and incorrect classifications across both outcome categories, thereby providing the foundation for calculating diagnostic metrics such as sensitivity, specificity, PPV, and NPV. Based on Table 4, Logistic regression after applying SMOTE demonstrated an overall accuracy of 75.2% on the independent test set.

Table 4: Confusion matrix for the Logistic regression classifier

	Observed	
Predicted	Positive	Negative
Positive	TP = 76	FP = 21
Negative	FN = 10	TN = 18

TP = PCR-positive patients correctly classified as positive;
TN = PCR-negative patients correctly classified as negative;
FN = PCR-positive patients incorrectly classified as negative;
FP = PCR-negative patients incorrectly classified as positive.

In the next experiment, we use the LDA method for classifying the COVID-19 data based on input variables. LDA classification leads to a linear function which is split the patients into two classes. The confusion matrix for the LDA model is shown in Table

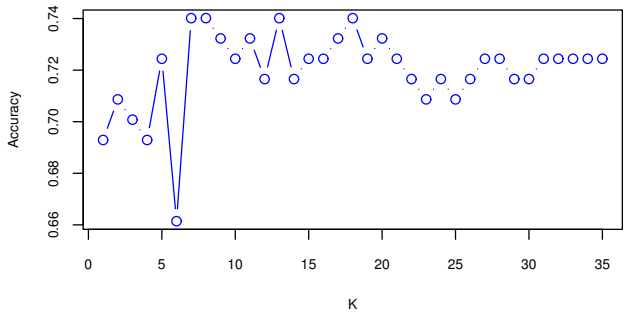

Figure 3: The accuracy of KNN models for different K's.

5. Based on Table 5, the accuracy of this model is 0.720. Thus, the Logistic regression is more accurate than the LDA model in the class of parametric classification methods.

Table 5: Confusion matrix for the LDA classifier

	Observed	
Predicted	Positive	Negative
Positive	67	15
Negative	19	24

The LDA model is a versatile method for the classification of the data. But, in this model, it is assumed that the input variables have a normal distribution. In our data, we have several categorical variables in the inputs. So, it should be careful that the normal assumption is not met in these data and that the LDA method causes to biased results.

Due to this limitation in the parametric methods, we also use some nonparametric models to compare them with the previous statistical learning methods. The KNN algorithm is a distance-based classification method which is not assumed any limitation in the distribution of input or output variables. As noted, the key parameter in the KNN method is k, which indicates the number of neighbors used in the KNN classifier. To select an optimal value for k, a 10-fold cross-validation approach is applied. The goal is to find a k which is maximized the accuracy of the KNN algorithm in an iterative bootstrap sampling scheme. Figure 3 shows the result of this method to choose the best value for k. It can be seen that a k=7 is maximized the accuracy of the KNN classifier. We apply this KNN classifier with k=7 to the testing dataset and obtain the confusion matrix (Table 6).

Another machine learning algorithm used in this study is DT. Figure 4 shows the prune decision tree for COVID-19 patients. The method of pruning the tree is discussed in Table 1. According to Figure 4, WBC and Immunodeficiency are two important variables in indicating the classes of each patient based on this algorithm. The terminal nodes of this tree are specified how many patients are categorized in each

Table 6: Confusion matrix for the KNN classifier

	Observed	
Predicted	Positive	Negative
Positive	60	14
Negative	26	25

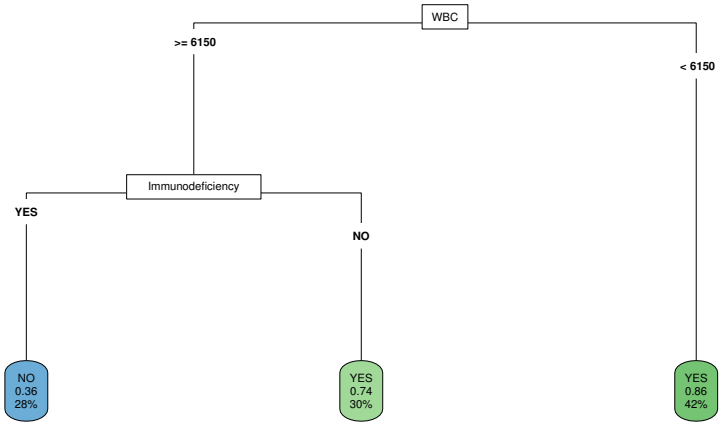

Figure 4: The prune decision tree for the Covid-19 patients.

class. For example, if for a patient, the value of WBC is less than 6150 then this patient is classified in the "YES" group (Positive COVID-19) with a probability of 0.86, and 42 percent of the train observations are fallen in this node. The other nodes can be interpreted in a similar way. The confusion matrix of the DT algorithm can be obtained like the previous methods which is shown in Table 7.

Table 7: Confusion matrix for the DT classifier

	Observed	
Predicted	Positive	Negative
Positive	72	19
Negative	14	20

The RF algorithm is the next machine learning method that is employed to diagnose COVID-19 patients. The RF is achieved by aggregating several DTs (ntree) each of which is grown using a random set of input variables (mtry). In order to choose the best values for ntree and mtry, we perform a cross-validation scheme by a grid search algorithm that is summarized in Table 1. Figure 5 shows the accuracy of random forests growing via several values of mtry and ntrees. According to this figure, an RF with mtry=6 and ntree=150 has a good performance in comparison with other choices. Table 8 shows the confusion matrix of the RF method.

Figure 6 displays the variable importance plot from the RF model. Among the pre-

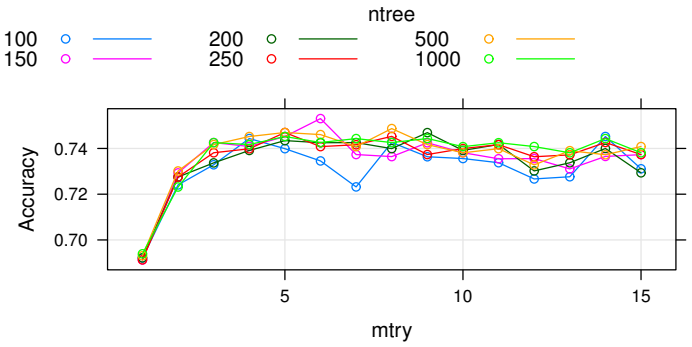

Figure 5: The accuracy of RF models for different mtry and ntrees.

Table 8: Confusion matrix for the RF classifier

	Observed	
Predicted	Positive	Negative
Positive	84	28
Negative	2	11

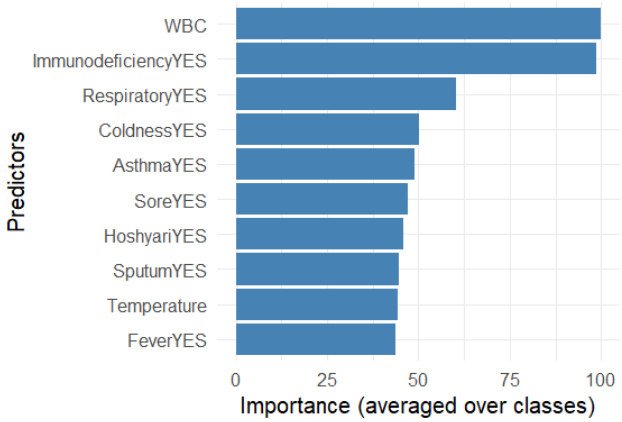

Figure 6: Bar plot for the rank of importance of the variables in RF model. (Only 10 most important variables are shown)

dictors, WBC and immunodeficiency showed the highest importance in discriminating PCR-positive from PCR-negative cases, highlighting their central role in the classification task. In contrast, fever had the lowest importance, suggesting that although clinically relevant, it contributed less to the predictive power of the model compared to laboratory and comorbidity-related features. Table 9 compares the performance of the five classifiers under two imbalance handling strategies (with and without SMOTE). Without SMOTE, RF model achieved the highest overall accuracy (76.8%) and balanced sensitivity (78.8%) and specificity (69.2%), whereas Logistic regression also performed

competitively (accuracy = 75.2%, AUC = 0.788). KNN and decision tree showed lower accuracy and AUC values. After applying SMOTE, sensitivity substantially improved across most models, with the largest gain observed for random forest (sensitivity = 97.6%) and Logistic regression (sensitivity = 88.3%). However, this improvement was accompanied by a notable decline in specificity (e.g., 28.2% for RF). Positive predictive value (PPV) remained high across models, while negative predictive value (NPV) improved under SMOTE, especially for Logistic regression (NPV = 64.2%) and RF (NPV = 84.6%). These results indicate that SMOTE effectively increases the ability to detect positive cases (higher sensitivity and NPV) but at the cost of misclassifying more negative cases, reducing specificity.

Table 9: Performance comparison of the classifiers in two imbalance handling methods

Imbalance handling	Classifier	Accuracy	Sensitivity	Specificity	PPV	NPV	AUC
Without SMOTE	Logistic regression	0.752	0.784	0.643	0.884	0.462	0.788
	LDA	0.736	0.757	0.636	0.907	0.359	**0.815**
	KNN	0.704	0.709	0.625	**0.965**	0.128	0.748
	DT	0.696	0.740	0.520	0.860	0.333	0.607
	RF	**0.768**	**0.788**	0.692	0.907	0.462	0.808
With SMOTE	Logistic regression	0.752	0.883	0.461	0.783	0.642	0.789
	LDA	0.720	0.779	0.615	0.817	0.558	0.804
	KNN	0.680	0.697	0.641	0.810	0.490	0.735
	DT	0.736	0.837	0.512	0.791	0.588	0.638
	RF	**0.760**	**0.976**	0.282	0.750	**0.846**	**0.816**

Concurrently, it is vital to acknowledge the appeal of machine learning methods, exemplified by RF, which exhibit robust performance without the need for these stringent assumptions. The outcome of this study thus positions RF as a preferred choice for COVID-19 patient classification. In addition, the incorporation of Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) metric are presented for all the classification models in Figure 7. The inclusion of Receiver Operating Characteristic (ROC) curve analysis and the assessment of the Area Under the Curve (AUC) metric offer valuable insights into the performance of these models. These metrics play a pivotal role in evaluating the discriminatory power of the models and their ability to strike a balance between sensitivity and specificity. ROC curves visually represent the trade-off between the sensitivity and the specificity at various classification thresholds. The curve showcases the model's ability to distinguish between COVID-19 patients and non-COVID-19 cases across different decision boundaries. A steeper ROC curve signifies better performance as it implies higher sensitivity without a proportionate increase in false positives. The AUC value, on the other hand, quantifies the overall performance of the model. It represents the area under the ROC curve, with a perfect model achieving an AUC of 1.0, while a random or non-discriminatory model scores an AUC of 0.5. An AUC greater than 0.5 indicates that the model has a higher likelihood of correctly classifying COVID-19 patients compared to random chance.

Figure 7 further substantiates the dominance of RF as a diagnostic tool for COVID-19. It is important to emphasize that RF consistently yielded higher AUC values and steeper ROC curves, indicating superior discriminatory power. While LDA demon-

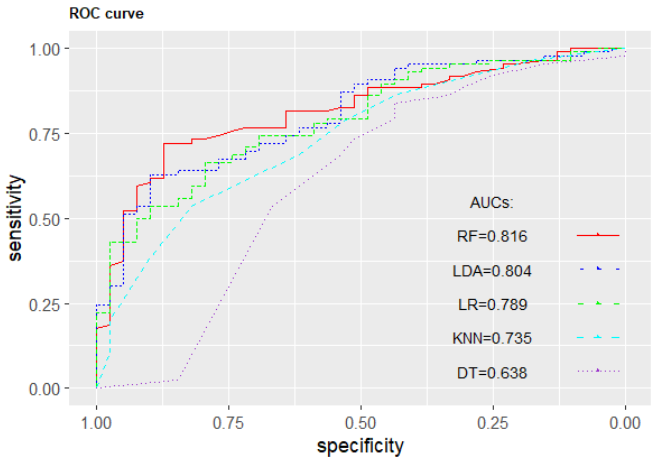

Figure 7: ROC curves for different classification methods with area under the curve (AUC) values of each of them.

strated commendable results, achieving competitive AUC values, the RF model consistently outperformed it across various evaluation metrics.

4 Discussion

Various methods are used to diagnose the COVID-19 disease, including examination of clinical symptoms related to the disease and diagnostic methods with higher accuracy such as PCR test, rapid test, and CT scan imaging of the lung (Li and Lu, 2020; Tang and Stratton, 2020). The rapid increase in patients and the importance of correct disease diagnosis have led to an increase in the desire to use high-precision and high-speed methods in order to quickly and accurately analyze the huge amount of patients' data and to make faster and more accurate decisions for doctors (Esfandiari and Tabar, 2014). In the present study, classification algorithms of predictors were used to diagnose the COVID-19 disease. New research studies demonstrate the integration of artificial intelligence (AI) and machine learning techniques to diagnose COVID-19 using medical imaging, specifically chest radiographs and CT scans. Artificial intelligence algorithms analyze imaging data to identify disease patterns, enabling rapid and accurate diagnosis (Butt and Babu, 2020). These approaches can help health care providers speed up diagnosis, especially when resources are limited or when disease cases are increasing (Wynants, 2020).

In the present study, five classification algorithms—LR, LDA, KNN, DT, and RF—were compared for the diagnosis of COVID-19 infection using demographic, clinical, and laboratory data. Before applying imbalance correction, RF and LR achieved the highest overall accuracies (76.8% and 75.2%, respectively) and AUC values above 0.78, demonstrating superior discriminatory ability compared to the other models. DT and KNN exhibited lower performance, likely due to their higher sensitivity to data imbalance

and noise.

After addressing class imbalance through the SMOTE technique, a marked improvement in sensitivity was observed across all models, particularly for RF (97.6%) and LR (88.3%), reflecting a stronger ability to identify COVID-19–positive cases. However, this gain was accompanied by a decrease in specificity (e.g., 28.2% for RF), indicating an increased tendency to classify samples as positive. The positive predictive value (PPV) remained high for most classifiers, while the negative predictive value (NPV) improved notably for LR (64.2%) and RF (84.6%). These findings suggest that SMOTE effectively enhances the detection of positive cases but at the expense of more false-positive classifications. From a clinical perspective, this trade-off may be acceptable in early diagnostic settings where missing true positives is more critical than false alarms.

The superior performance of the RF and LR models in this study aligns with previous research that has demonstrated the robustness of ensemble and regression-based classifiers in medical diagnostics. For instance, Mohammadi (2021) reported accuracies above 98% for artificial neural network (ANN) and LR models in COVID-19 diagnosis, while Zakariaee and Kazemi-Arpanahi (2023) found that RF achieved 97.2% accuracy with nearly perfect sensitivity and specificity in predicting patient mortality. Similar findings have been described by Shafqat and Ahmad (2020) and Schober and Vetter (2021), who highlighted the balance between interpretability and predictive power in these models. However, unlike prior studies that focused primarily on model performance under balanced or idealized data conditions, the present study explicitly addresses the class imbalance problem inherent in real-world clinical datasets by incorporating the SMOTE. This approach provides a more realistic and statistically rigorous comparison of classifiers, emphasizing not only accuracy but also sensitivity, specificity, and predictive values under both imbalanced and balanced data scenarios.

Moreover, in the present study, the random forest model demonstrated particularly stable and consistent performance, which can be attributed both to its intrinsic robustness and to the comprehensive preprocessing applied to the dataset. Missing and noisy values were addressed during data preparation, and outliers were carefully examined to minimize their influence on model training. This methodological rigor, combined with the random forest's internal mechanisms for handling residual noise and missing information, contributed to the model's resilience when applied to real clinical data Chahar and Roy (2022); AlJame and Mohammed (2021); Xia and Zhou (2023). In addition, the algorithm's ability to quantify variable importance enhanced interpretability, revealing key predictors—such as white blood cell count (WBC) and immunodeficiency—as the most influential factors distinguishing PCR-positive from PCR-negative patients Iwendi (2020).

Overall, the findings indicate that both LR and RF are effective for classifying COVID-19 test outcomes, with RF offering higher sensitivity and LR providing clearer interpretability. The incorporation of SMOTE improved sensitivity and NPV, underscoring the importance of addressing data imbalance in medical prediction tasks. Nevertheless, reduced specificity after oversampling highlights the need for threshold calibration and external validation before clinical implementation. These insights reinforce the potential of machine learning, when properly optimized and interpreted, to

support clinical decision-making and early identification of infectious diseases such as COVID-19.

5 Conclusion

Different modeling methods, such as logistic regression and random forest, promise a transformation in the diagnosis and management of the COVID-19 disease. According to the results of the present study and the ability of these models to manage complex data sets, identify patterns, and predict results, they can be used to help improve and accelerate the clinical decision-making process and improve patient care. Moreover, these models can assist clinicians in the early identification of patients likely to test positive for COVID-19 at admission, which may help prioritize confirmatory testing and resource allocation. Further research is required to determine whether their use affects treatment decisions or patient survival.

Acknowledgements

We thank the anonymous reviewers and the editorial board for their valuable comments. This research was financially supported by Rafsanjan University of Medical Sciences (Project No. 99350).

References

AlJame IAAI M, Mohammed A. Deep forest model for diagnosing COVID-19 from routine blood tests. Scientific reports. 2021;11(1):16682.

Butt GJCD C, Babu BA. Deep learning system to screen coronavirus disease 2019 pneumonia. Appl Intell. 2020;p. 16682.

Chahar S, Roy PK. Covid-19: A comprehensive review of learning models. Archives of Computational Methods in Engineering. 2022;29(3):1915–1940.

Chawla BKWHLO N V, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research. 2002;16:321–357.

Esfandiari BMRMAME N, Tabar VK. Knowledge discovery in medicine: Current issue and future trend. Expert Systems with Applications. 2014;41(9):4434–4463.

Fischhoff B. Making decisions in a COVID-19 world. JAMA. 2020;324(2):139–140.

Iwendi BAKPASRCJMPSJO C. COVID-19 patient health prediction using boosted random forest algorithm. Frontiers in public health. 2020;8:357.

James WDHT G, Tibshirani R. An introduction to statistical learning. Springer; 2013.

Johnson RA, Wichern DW. Applied multivariate statistical analysis. Wiley; 2002.

Li GMPYML X, Lu S. Molecular immune pathogenesis and diagnosis of COVID-19. Journal of pharmaceutical analysis. 2020;10(2):102–108.

Malik D, Munjal G. Reviewing classification methods on health care. Intelligent Healthcare: Applications of AI in eHealth. 2021;1:127–142.

Mohammadi PHKHMMMMANSM F. Artificial neural network and logistic regression modelling to characterize COVID-19 infected patients in local areas of Iran. Biomedical journal. 2021;44(3):304–316.

Organization WH, Statistics NCfH. The International Classification of Diseases, 9th Revision, Clinical Modification: Procedures: tabular list and alphabetic index, vol. 3. Commission on Professional and Hospital Activities.; 1980.

Rashidi ASRSTNK H H, Hu B. Common statistical concepts in the supervised Machine Learning arena. Frontiers in Oncology. 2023;13:1130229.

Schober P, Vetter TR. Logistic regression in medical research. Anesthesia and Analgesia. 2021;132(2):365–366.

Shafqat KSRRUQJAT S, Ahmad HF. Big data analytics enhanced healthcare systems: a review. The Journal of Supercomputing. 2020;76:1754–1799.

Sidey-Gibbons JA, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. BMC medical research methodology. 2019;19:1–18.

Tang SJEPDH Y W, Stratton CW. Laboratory diagnosis of COVID-19: current issues and challenges. Journal of clinical microbiology. 2020;58(6):512–520.

Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association. 2018;113(523):1228–1242.

Weisberg S. Applied linear regression. Wiley; 2005.

Wynants VCBCGSRRDHGSEBMM L. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ. 2020;369.

Xia ZP Y, Zhou Z. Analysis And Prediction of COVID-19 Based on Machine Learning. Highlights in Science, Engineering and Technology. 2023;38:725–735.

Zakariaee NNEM S S, Kazemi-Arpanahi H. Comparing machine learning algorithms to predict COVID-19 mortality using a dataset including chest computed tomography severity score data. Scientific reports. 2023;13(1):11343.