scientific reports

OPEN

Comparative analysis of feature selection techniques for COVID-19 dataset

Farideh Mohtasham[1]✉, MohamadAmin Pourhoseingholi[2], Seyed Saeed Hashemi Nazari[3], Kaveh Kavousi[4]✉ & Mohammad Reza Zali[1]

In the context of early disease detection, machine learning (ML) has emerged as a vital tool. Feature selection (FS) algorithms play a crucial role in ensuring the accuracy of predictive models by identifying the most influential variables. This study, focusing on a retrospective cohort of 4778 COVID-19 patients from Iran, explores the performance of various FS methods, including filter, embedded, and hybrid approaches, in predicting mortality outcomes. The researchers leveraged 115 routine clinical, laboratory, and demographic features and employed 13 ML models to assess the effectiveness of these FS methods based on classification accuracy, predictive accuracy, and statistical tests. The results indicate that a Hybrid Boruta-VI model combined with the Random Forest algorithm demonstrated superior performance, achieving an accuracy of 0.89, an F1 score of 0.76, and an AUC value of 0.95 on test data. Key variables identified as important predictors of adverse outcomes include age, oxygen saturation levels, albumin levels, neutrophil counts, platelet levels, and markers of kidney function. These findings highlight the potential of advanced FS techniques and ML models in enhancing early disease detection and informing clinical decision-making.

Machine Learning (ML) has been widely used in building classification models for the early detection of diseases using electronic health records (EHRs) of patients. However, there has been an increasing number of challenges and complexities in building accurate classification models. Feature Selection (FS) is crucial for handling large or high-dimensional data like hospital records, as it identifies the most relevant features, enabling the creation of a low-density dataset[1,2].

Filter, Wrapper, and embedded methods are three groups of FS methods[3]:

- Filter methods use feature ranking as the evaluation metric for selecting features.
- Wrapper methods use the performance of a chosen classifier algorithm to help select the best feature subset while considering the interdependencies among features.
- Embedded methods use weights or importance given to each feature to produce the best classification accuracy, which is determined during the training step.
- Hybrid feature selection methods integrate the principles of filter, wrapper, and embedded approaches. They utilize a multi-step process that first applies filter methods for initial feature reduction based on statistical measures, followed by wrapper or embedded methods for more nuanced selection, capitalizing on the computational efficiency of filter methods and the predictive accuracy of wrapper and embedded techniques[4,5].

Various FS methods have been used to predict COVID-19 clinical outcomes, but there's a lack of comprehensive analysis on the different methods.

Ali and Abdulsalam[6] used two FS methods, Recursive Feature Elimination and Extra Tree Classifier (ETC), to determine significant features. They applied Naïve Bayesian and Restricted Boltzmann Machine classification

[1]Gastroenterology and Liver Diseases Research Center, Research Institute for Gastroenterology and Liver Diseases, Shahid Beheshti University of Medical Sciences, Tehran, Iran. [2]Hearing Sciences, Mental Health and Clinical Neurosciences, School of Medicine, National Institute for Health and Care Research (NIHR) Nottingham Biomedical Research Center, University of Nottingham, Nottingham, UK. [3]Department of Epidemiology, School of Public Health & Safety, Shahid Beheshti University of Medical Sciences (SBMU), Tehran, Iran. [4]Laboratory of Complex Biological Systems and Bioinformatics (CBB), Department of Bioinformatics, Institute of Biochemistry and Biophysics (IBB), University of Tehran, Tehran, Iran. ✉email: f-mohtasham@sbmu.ac.ir; kkavousi@ut.ac.ir

methods, and found that using the top ten features selected by the ETC method improved classification accuracy to 66.33% and 99.92%.

Pourhomayoun and Shakibi[7] applied different filter and wrapper methods to select the best feature subset, without explicitly mentioning the names of the methods or conducting a comparison between them.

Varzaneh et al.[8] compared different meta-heuristic algorithms for FS and found the Horse herd Optimization Algorithm (HOA) to be the most effective based on three performance evaluation metrics, mean fitness value, classification accuracy using K-Nearest Neighbors (KNN), and the number of selected features.

Hayet-Otero et al.[9] conducted an extensive analysis of 166 FS methods, evaluating stability, similarity, computation time, and number of selected features. The study did not address class imbalance or explore hybrid methods, and achieved classification accuracy through a limited set of ML algorithms.

Imbalanced training data presents a significant challenge in predictive modeling[10], especially in real-world scenarios such as COVID-19[11].

In response, our study employs filter, embedded, and hybrid FS methods on a balanced dataset. We assess the effectiveness of these methods using 13 ML models, evaluating based on classification accuracy, predictive accuracy, similarity, and the number of selected features. This approach aims to address gaps in existing literature by providing a comprehensive comparison and demonstrating the benefits of hybrid methods in managing high-dimensional data.

Materials and Methods

The sources of data for the development and evaluation were a dataset consisted of EHRs of COVID-19 patients admitted to three educational hospitals in Tehran, Iran from March 20th, 2020 to March 18th, 2021. They were obtained through a formal request process. Institutional approvals were obtained from the hospitals' ethics committees, and all ethical considerations were taken into account during the data collection process. Patient confidentiality and privacy were strictly maintained, and all necessary permissions were obtained before accessing and using the EHRs for research purposes.

The decision to utilize this dataset stemmed from its direct relevance to the pressing global health issue of COVID-19 and the critical need for accurate predictive models to aid in healthcare decision-making. The dataset, comprising detailed clinical, laboratory, and demographic information from 4778 COVID-19 patients in Iran, offers a rich source of varied features that are essential for developing robust prediction models. Moreover, the study conducted by Hatamabadi et al.[12] offers a comprehensive exploration of the cohort's epidemiology, enriching the understanding of the dataset's context.

Regarding the representativeness of the data, the dataset captures a diverse patient population affected by COVID-19 within a specific geographical context. By including a wide range of clinical and demographic variables, the dataset is believed to provide a comprehensive representation of factors influencing COVID-19 mortality outcomes in this population. Therefore, the utilization of this dataset is essential in generating insights and models that can potentially be generalized to similar patient cohorts and contribute to improved clinical decision-making in the context of COVID-19.

Step 1: Data preprocesssing

Robust scaling[13] was employed to normalize the data and mitigate the impact of outliers. This method, ideal for datasets with non-normal distributions or extreme values, normalizes data by removing the median and scaling based on the interquartile range, thereby reducing sensitivity to outliers and extreme values, making it a more reliable method for normalizing data and reducing the impact of outliers on the analysis. This approach is beneficial because it allows for more accurate and consistent results, particularly in datasets with non-normal distributions or significant outliers.

For achieving transportability, a random stratified split was employed, allocating 70% for training and 30% for testing. The development dataset comprised 3345 cases, while 1433 cases were reserved for model evaluation.

A multicenter, cross-sectional study was conducted on COVID-19 cases hospitalized in 19 public hospitals affiliated with Shahid Beheshti University of Medical Sciences (SBMU) in Tehran, Iran, between February 19 and May 12, 2020[14]. The study revealed a case fatality rate (CFR) of 10.05% from which we inferred a 10% likelihood of experiencing the Death outcome. In our modeling process, we plan to incorporate a maximum of 20 predictor variables. It is recommended that models with rare binary predictors have a minimum of 20 events per variable[15]. Therefore, recruiting 2000 patients (200/0.1) for the study is recommended.

Step 2: Selection of relevant features

The variance inflation factor (VIF) was used to measure collinearity among independent variables before FS[16,17], and a correlated feature was dropped to avoid overfitting and inaccurate performance.

Filter, embedded and hybrid FS methods applied to get the relevant features, which are.

Filter Methods

Conditional mutual information maximization (CMIM)

We chose the CMIM method due to its unique ability to select features based on their relevance to the target variable while simultaneously considering the redundancy between features[18]. By iteratively selecting features that maximize mutual information with the target variable while maintaining conditional independence from previously selected features, CMIM can effectively identify a compact subset of features that collectively capture the key predictive information without introducing redundancy[19,20].

The decision to utilize CMIM was driven by our objective to efficiently handle high-dimensional data and extract a concise yet informative set of features for predictive modeling. The method's capability to balance relevance, diversity, and non-redundancy in FS aligns with our goal of building a robust and interpretable model.

The CMIM filter method was implemented using the mlr3filters package[21] for its functionality, support for FS techniques, reproducibility, and transparency in the FS process, which aids in validating and replicating results by other researchers.

Correction Filter

We opted for the Correlation filter[22] method due to its simple and intuitive approach in identifying features strongly associated with the target variable. By calculating correlation coefficients for both numeric and categorical features, we aimed to capture linear and non-linear relationships beneficial for predictive modeling. The 0.2 threshold for feature selection was chosen to prioritize features showing moderate to strong correlation with the target variable, potentially enhancing our model's predictive accuracy. This method was chosen for its simplicity and interpretability, enabling a transparent FS process to improve the overall robustness of our analysis.

Embedded method: mean decrease Gini (MDG) based Random Forest

Mean Decrease Gini (MDG) is a FS method based on Random Forest algorithms[23,24]. It works by measuring the importance of each feature in the Random Forest model by calculating the decrease in Gini impurity that each feature causes when it is used to split the data. The Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset. The higher the decrease in Gini impurity, the more important the feature is considered to be[25].

MDG is chosen as a robust and reliable FS method that can handle high-dimensional data and capture non-linear relationships between features and the target variable.

Data preparation for Random Forest involved standardizing the data, creating a 70%/30% train/test split, addressing dataset imbalance, and balancing data. The optimal number of random variables (mtry) was determined by building 10 RF classifiers with varying numbers of trees (ntree) and identifying where the out-of-bag error rate stabilized. The top 20 important variables were identified using the mean decrease in node impurity method.

Hybrid Methods

Hybrid of ANOVA, backward selection and Lasso methods namely ABL

ABL is a hybrid method that fuse ANOVA, backward selection, and Lasso methods to select relevant features.

ANOVA is a parametric technique that could be used with a specific dataset feature to equate 'class means'[26]. It assesses statistically significant difference between class means for each feature using the F-test statistic and p-value. The premise of ANOVA is to find a score for each feature saying "how well this feature discriminates between classes" to remove the irrelevant and unwanted features[27].

Sequential backward elimination (SBE)[28] works by iteratively removing the least significant features from the model until a stopping criterion is met. It starts with all features included and then removes one feature at a time, re-evaluating the model's performance after each removal. The rationale behind SBE is to reduce the dimensionality of the feature space by eliminating features that contribute the least to the model's predictive power, which can lead to improved model interpretability and reduced overfitting.

Lasso regularization[29] works by adding a penalty term to the standard linear regression cost function, which penalizes the absolute size of the regression coefficients. This penalty encourages the regression coefficients of less important features to be reduced to zero, effectively performing FS by eliminating irrelevant features from the model. Lasso regularization is selected to simultaneously perform FS and regularization, in order to reduce overfitting and enhance model interpretability by encouraging sparsity in the feature space.

Hybrid of Boruta and combination of variable importance of multiple classification methods namely Boruta-VI

The Boruta-VI hybrid method combines the Boruta FS with variable importance from eleven classification models.

Boruta works by using a random forest algorithm to determine the importance of features. It compares the importance of real features with that of randomly shuffled, uninformative features. If a feature's importance is significantly higher than the maximum importance of the shuffled features, it is considered relevant. Boruta iteratively identifies and confirms relevant features, providing a robust and reliable FS method[30].

The reason for choosing Boruta is its ability to handle high-dimensional datasets and capture complex, non-linear relationships between features and the target variable.

Variable importance across different algorithms involves assessing the importance of features by using multiple ML algorithms and comparing their evaluations. By applying various algorithms, such as random forests, gradient boosting machines, or support vector machines, the method can capture different aspects of feature relevance and provide a more comprehensive understanding of feature importance.

This approach is chosen to gain a diverse perspective on feature importance, as different algorithms may emphasize different aspects of the data.

In hybrid methods, the features selected in each step are included in the subsequent step.

Step 3: Model Development and Model Specification

The selection of appropriate models was grounded in a meticulous review of pertinent literature. Bottino et al.[31] conducted an exhaustive analysis of machine learning methodologies for COVID mortality prediction, leading to the identification of key algorithms utilized in previous studies. The chosen models were based on their specific strengths and suitability for clinical applications, as outlined below:

- *Logistic Regression* Chosen for its efficiency in computations and interpretability, making it well-suited for clinical settings.
- *Support Vector Machine (SVM)* A non-linear statistical supervised learning tool.
- *K-Nearest Neighbors (KNN)* Acknowledged as a straightforward and longstanding method for pattern classification.
- *Naïve Bayes* A supervised algorithm founded on Bayes' theorem.
- *Decision Tree Classifier* A non-parametric supervised algorithm.
- *Extreme Gradient Boosting Algorithm* An ensemble method known for excellent performance in binary classification tasks due to its effective capture of nonlinearity and predictor interactions through a recursive tree-based decision framework.
- *Random Forest (RF)* An ensemble learning model characterized by multiple decision trees, esteemed as a leading algorithm in the field.

The specified learning algorithms along with their hyperparameters were:

1. Generalized Linear Model (GLM): Default settings.
2. Linear Discriminant Classifier (LDC): Default settings.
3. Regularized Regression (lasso): Hyperparameters selected as alpha = 1, lambda = 0.0014.
4. K-Nearest Neighbors (K-NN): Hyperparameter k = 1.
5. Naive Bayes: Default settings.
6. Support Vector Machine Classification (SVM): Tuned hyperparameters sigma = 0.15, C = 10.
7. Classification and Regression Trees (CART): Method 'rpart' with default settings.
8. Decision Tree (C5.0): Tuned hyperparameters trials = 50, model = tree, winnow = FALSE.
9. Random Forest (RF): Hyperparameters mtry = 3, ntree = 700.
10. Bagged CART (treebag): Default settings.
11. Stochastic Gradient Boosting (GBM): Hyperparameters n.trees = 250, interaction.depth = 5, shrinkage = 0.1, n.minobsinnode = 10.
12. Extreme Gradient Boosting (XGBOOST): Hyperparameters nrounds = 250, max_depth = 5, eta = 0.4, gamma = 0, colsample_bytree = 0.8, min_child_weight = 1, subsample = 1.
13. Neural Network (NN): Hyperparameters size = 10, decay = 0.1.

Once trained, the predictive performance of the models was rigorously assessed through a thorough 10 times repeated tenfold cross-validation[32] approach. This internal validation technique provided estimates of model performance on unseen data and allowed for fine-tuning of model parameters. The outcomes from cross-validation were utilized as a benchmark for constructing the final classification model and facilitating model selection. Subsequently, the final model underwent validation against an independent test dataset to confirm its predictive efficacy (external validation).

The caret package[33,34] was employed for model training, comparison, and subsequent analyses.

Step 4: Model Performance

In our study, we rigorously evaluated and compared machine learning algorithms using a comprehensive set of standard discrimination and calibration evaluation metrics.

Firstly, we assessed discrimination using metrics such as accuracy, sensitivity, specificity, precision, F1-score, Kappa, and ROC curves. These metrics help us understand how well the models distinguish between classes and make accurate predictions.

Secondly, for calibration evaluation, we utilized calibration plots to visualize the agreement between predicted probabilities and actual outcomes. This plot provides insights into the calibration of the models and helps in understanding their reliability in predicting the probabilities of events.

To compare the performance of different feature selection methods and models, we conducted statistical tests to identify significant differences in their predictive capabilities. By analyzing the results of these tests, we gained valuable insights into the relative strengths and weaknesses of each approach.

Overall, our evaluation framework incorporated a diverse range of measures and plots to comprehensively assess model performance, ensure calibration, and facilitate meaningful comparisons among various methodologies.

Ethical approval

All methods were performed in accordance to Helsinki protocol. The Institutional Review Board (IRB) at the Shahid Beheshti University of Medical Science approved the study and waived informed consent gathering (IR.SBMU.RIGLD.REC.1401.032). Data were anonymized before analysis, and patient confidentiality and data security were concerned.

Results

4778 COVID-19 patients were included in the study, which examined 116 routine clinical, laboratory, and demographic features. The fatality rate in this cohort was 22% (N = 1050, 59.6% male). The mean ± SD age for the death and surviving groups were 70.8 ± 15.6 and 58.3 ± 16.9 years, respectively.

After splitting the data and creating a 70%/30% train/test split, we discovered that in our training dataset, 22% of instances belong to the "Death" class and 78% are labeled as "alive". This is significant because it indicates a moderate imbalance in the data. To address this, we utilized rebalancing techniques to prevent the minority class from being disproportionately represented in the training set[35] before applying machine learning algorithms. We then used the balanced dataset to make predictions. Using the ROSE package[36], we implemented both over-sampling of the minority class with replacement and under sampling of the majority class without replacement to balance the distribution while maintaining the integrity of the dataset. The resulting data had the same size as the original dataset with a 1:1 ratio of "Death" to "Alive".

Number of selected features in every FS method

Using the VIF, the dataset's features were reduced to 109 predictors. The CMIM approach identified six key features, whereas the correlation filter method discerned significant correlations for eight numerical and strong Phi Correlation Coefficients for three categorical features.

The Gini impurity embedded method, utilizing 700 trees, with five random variables each, selected twenty important features, achieving a minimum out-of-bag error rate of 5.29%.

The Hybrid ABL method eliminated 39 features based on ANOVA results, and further feature reduction through backward selection and Lasso resulted in 15 predicting features.

Using the Boruta-VI hybrid method identified 55 attributes were initially deemed significant, which were then narrowed down to 20 based on importance scores across eleven prediction models (Supplementary Tables S1–S4, Supplementary Figs. S1, S2).

Comparison of selected features

The Jaccard index[37] was utilized to compute the similarity between pairs of FS methods. As anticipated, the Jaccard index for Hybrid Boruta-VI—MDG Random Forest (0.6) and Correlation–CMIM (0.54) is high among the combinations, while Hybrid ABL demonstrates low similarity with other FS methods (Fig. 1).

The most important features that included in every FS method were age and neutrophil count (NEUT), followed by oxygen saturation (O2sat), Albumin, UREA, and blood urea nitrogen (BUN) (Table 1).

To investigate the relationship between the occurrence of Death and the predictor included in each FS method, we employed multivariable binary logistic regression analysis. In the Hybrid ABL dataset, at a significance level of 5%, all predictors with the exception of CKD, FBS, and ESR were found to be significantly correlated with Death (p-values < 0.05). Similarly, in the correlated dataset, at the 5% significance level, all predictors except LYMPHH, BUN, and CR were determined to be significantly associated with Death (p-values < 0.05). In the CMIM dataset, all predictors, except BUN, were identified as significantly associated with Death at the 5% significance level. For the MGD dataset, at the 5% significance level, all predictors, excluding WBC, CR, LDH, PT, HB, and TIBC,

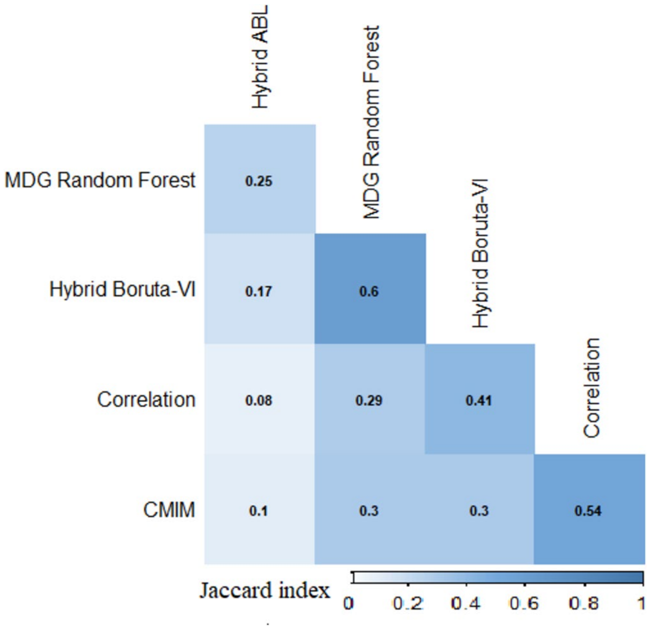

Figure 1. The similarity of the features selected per each feature selection method.

CMIM (6 features)	Correlation (11 features)	Hybrid ABL (15 featured)	MGD Random forest (20 features)	Hybrid Boruta-VI (20 features)
Age	Age	Age	Age	Age
NEUT	NEUT	NEUT	NEUT	NEUT
O2sat	O2sat		O2sat	O2sat
ALBUMIN	ALBUMIN		ALBUMIN	ALBUMIN
UREA	UREA		UREA	UREA
BUN	BUN		BUN	BUN
	CR		CR	CR
		Lactate dehydrogenase (LDH)	Lactate dehydrogenase (LDH)	Lactate dehydrogenase (LDH)
		Ferritin	Ferritin	Ferritin
		P	P	P
	Decreased Consciousness			Decreased Consciousness
	Dialysis			Dialysis
		O2sat.Ventilator	O2sat.Ventilator	
		TIBC	TIBC	
			WBC	WBC
			CR	CR
			LDH	LDH
			PROBNP	PROBNP
			INR	INR
			BE	BE
	LYMPHH			
	Blood.Injection			
		FBS		
		Procalcitonin		
		Serum Sodium		
		Muscle.Pain		
		Chronic kidney disease (CKD) Comorbidity		
		Taste or Smell loss (Taste.Smell)		
		D-dimer		
		Erythrocyte sedimentation rate (ESR)		
			PT	
			HB	
			LDL	
				Platelets.FFP.Injection
				PLT
				FDP

Table 1. Selected features in every FS method.

were deemed significantly associated with Death. Lastly, in the Boruta-VI dataset, at the 5% significance level, all predictors except CR and UREA were found to be significantly associated with Death (p-values < 0.05) (Table 2).

In comparing the binary logistic regression models for each FS dataset using the anova() function, significant differences in predictive performance with respect to Death were observed at the 5% significance level. Specifically, the Hybrid Boruta-VI and MGD models were found to be significantly superior to the CMIM, correlated, and Hybrid ABL models in predicting Death (Pr(> Chi): 2.2e−16).

On the other hand, the CMIM and Hybrid ABL models (Pr(> Chi): 26.525), the correlated model and the Hybrid ABL models (Pr(> Chi): – 159.16) were deemed to be statistically equivalent in predictive performance for Death. Additionally, the Hybrid Boruta-VI and MGD models also showed no significant difference in predictive performance (Pr(> Chi): – 124.74) for Death at the 5% significance level.

To establish relationships between the selected features and the outcome class and to explore potential interactions among these features, a graphical representation of the correlation matrix was generated for the Hybrid Boruta-VI dataset utilizing the corrplot package[38]. This visualization highlighted the most correlated variables, with correlation coefficients represented by color gradients. Positive correlations were depicted in shades of blue, while negative correlations were visualized in varying intensities of red. The intensity of the color and the size of the circles were indicative of the strength of the correlation coefficients, providing a visually intuitive

	Characteristic	OR	95% CI	p-value
Hybrid ABL	Muscle.Pain Yes	0.61	0.50, 0.74	< 0.001
	Taste.Smell Yes	0.22	0.07, 0.62	0.007
	CKD Yes	1.33	0.94, 1.89	0.11
	Age	3.56	3.11, 4.09	< 0.001
	O2sat.Ventilator	0.70	0.65, 0.76	< 0.001
	NEUT	1.52	1.36, 1.70	< 0.001
	NA	1.12	1.02, 1.23	0.014
	P	1.29	1.19, 1.39	< 0.001
	FBS	1.07	0.98, 1.17	0.2
	ESR	0.93	0.82, 1.06	0.3
	LACTATE	1.26	1.17, 1.35	< 0.001
	PROCALCITONIN	1.02	1.01, 1.04	< 0.001
	Ferritin	1.56	1.40, 1.74	< 0.001
	TIBC	0.87	0.77, 0.97	0.017
	DDIMER	1.07	1.01, 1.14	0.020
Correlated	Decreased.Consciousness Yes	2.36	1.83, 3.05	< 0.001
	Dialysis Yes	3.11	2.09, 4.71	< 0.001
	Blood.Injection Yes	2.00	1.57, 2.56	< 0.001
	Age	2.90	2.51, 3.36	< 0.001
	O2sat	0.65	0.60, 0.69	< 0.001
	LYMPHH	1.15	0.85, 1.55	0.4
	NEUT	1.63	1.21, 2.19	0.001
	BUN	0.99	0.92, 1.07	0.8
	CR	0.98	0.94, 1.03	0.5
	ALBUMIN	0.59	0.53, 0.65	< 0.001
	UREA	1.25	1.13, 1.38	< 0.001
CMIM	Age	2.87	2.50, 3.30	< 0.001
	O2sat	0.65	0.61, 0.69	< 0.001
	NEUT	1.45	1.30, 1.62	< 0.001
	BUN	1.01	0.94, 1.08	0.9
	ALBUMIN	0.56	0.50, 0.61	< 0.001
	UREA	1.35	1.25, 1.47	< 0.001
MGD	Age	3.10	2.66, 3.62	< 0.001
	O2sat	0.70	0.65, 0.76	< 0.001
	O2sat.Ventilator	0.90	0.81, 1.00	0.048
	WBC	1.06	0.98, 1.15	0.15
	NEUT	1.35	1.20, 1.52	< 0.001
	HB	1.00	0.89, 1.12	> 0.9
	BUN	0.90	0.82, 0.98	0.015
	CR	1.02	0.97, 1.07	0.5
	P	1.21	1.10, 1.33	< 0.001
	LDL	0.70	0.62, 0.79	< 0.001
	ALBUMIN	0.62	0.55, 0.69	< 0.001
	LDH	1.08	0.99, 1.17	0.079
	LACTATE	1.26	1.18, 1.36	< 0.001
	PROBNP	1.12	1.06, 1.19	< 0.001
	PT	1.00	0.90, 1.06	> 0.9
	INR	1.11	1.03, 1.25	0.014
	BE	0.76	0.69, 0.85	< 0.001
	Ferritin	1.39	1.24, 1.56	< 0.001
	TIBC	1.02	0.91, 1.16	0.7
	UREA	1.12	1.01, 1.24	0.039
Continued				

	Characteristic	OR	95% CI	p-value
Hybrid Boruta-VI	Decreased.Consciousness Yes	2.37	1.83, 3.10	<0.001
	Dialysis Yes	2.98	1.91, 4.71	<0.001
	Platelets.FFP.Injection Yes	4.64	2.82, 7.91	<0.001
	Age	3.09	2.65, 3.61	<0.001
	O2sat	0.67	0.62, 0.72	<0.001
	WBC	1.10	1.01, 1.21	0.038
	NEUT	1.34	1.19, 1.51	<0.001
	PLT	0.79	0.71, 0.87	<0.001
	BUN	0.91	0.84, 1.00	0.047
	CR	0.96	0.92, 1.02	0.2
	P	1.22	1.11, 1.34	<0.001
	ALBUMIN	0.67	0.60, 0.75	<0.001
	LDH	1.10	1.02, 1.20	0.020
	LACTATE	1.16	1.08, 1.25	<0.001
	PROBNP	1.12	1.06, 1.18	<0.001
	INR	1.09	1.03, 1.16	0.003
	BE	0.78	0.70, 0.87	<0.001
	FDP	1.12	1.05, 1.21	0.001
	Ferritin	1.45	1.29, 1.62	<0.001
	UREA	1.08	0.97, 1.21	0.2

Table 2. The odds of Death for each predictor included in each feature selection method. *OR* odds ratio, *CI* confidence interval.

representation of the relationship between variables. It is pertinent to note that correlations with a p-value exceeding 0.01 were considered statistically insignificant and were included in the graphical representation.

Of particular significance within the correlation matrix is the robust positive correlation of 0.63 observed between UREA and CR, indicative of a strong positive relationship between these parameters. Additionally, the moderate positive correlation coefficient of 0.48 between UREA and BUN suggests a moderate positive association between these variables. These observed correlations substantiate the close interrelation among UREA, BUN, and CR, underlining their relevance in assessing kidney function within the dataset. Furthermore, the moderate positive correlations between UREA and both P (0.42) and PROBNP (0.30) unveil potential connections between UREA levels and phosphate levels as well as heart function, considering PROBNP as a recognized marker for cardiac strain. These findings illuminate possible interdependencies among these variables, hinting at intricate physiological relationships that could offer valuable insights into the underlying mechanisms influencing the observed dataset trends (Fig. 2).

Evaluation of classification models' performance in different feature selection method

Figure 3 and Supplementary Table S5 show a rigorous evaluation of the model performance estimation using 10 repeated tenfold cross validation for each algorithm across various feature selection methods. The presentation of accuracy outputs, along with confidence intervals, showcases that the Random Forest algorithm consistently outperformed other algorithms in terms of accuracy across all FS methods.

To further delve into the model's predictive powers, ROC curves were generated to summarize the trade-of between the true positive and false positive rates. Random Forest, C5.0, GBM, XGBOOST and Bagged CART models exhibited ROC curves that trended towards the upper left corner, indicating optimal sensitivity and specificity levels across different probability thresholds within our resampling method (Fig. 4).

Calibration plots focusing on the Random Forest algorithm (our best predictive model) across various feature selection methods (Fig. 5) indicated minimal bias, with point estimates closely centered on zero for each feature selection method (the calibration intercepts obtained from the val.prob function within the rms library[5] in R).

The model's performance was further validated through external validation using independent test datasets distinct from the training database. Table 3 indicated that the predictive performance of the Classification and Regression Trees (CART) model was notably low, whereas the Random Forest model demonstrated superior performance across all feature selection methods. Additionally, Fig. 6 highlighted the consistent excellence in discrimination, calibration, and predictive performance of the Random Forest model across all feature selection methods.

When comparing the F1-score, Accuracy, Area Under the Curve (AUC), and precision of various machine learning models across different feature selection methods in Fig. 7, it is evident that Hybrid Boruta-VI exhibits competitive performance but does not consistently outperform other methods across all evaluated metrics and feature selection strategies. Specifically, Hybrid Boruta-VI demonstrates a commendable F1-score, although comparable or marginally superior scores are achieved by other feature selection methods such as MDG and the absence of feature selection (Without FS). In terms of accuracy, while Hybrid Boruta-VI performs well, similar

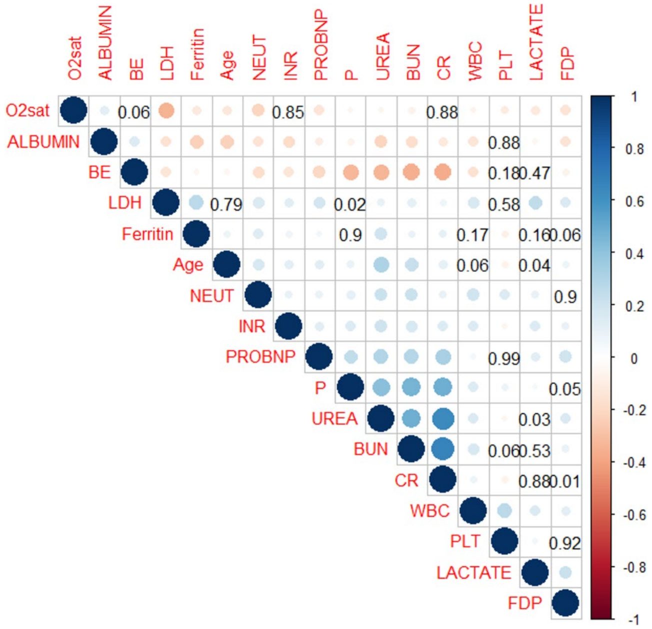

Figure 2. Correlation between the selected features and the outcome class in the Hybrid Boruta-VI dataset.

or higher levels of accuracy are observed with methods like Correlation and Without FS. Moreover, the AUC values obtained utilizing Hybrid Boruta-VI are not the highest among the compared feature selection techniques; instead, methods like Correlation and Without FS showcase superior AUC scores. Regarding precision, Hybrid Boruta-VI exhibits good results, yet alternative methods like Correlation and Without FS present comparable or enhanced precision metrics.

In light of these nuanced findings, statistical analyses were conducted to thoroughly evaluate the efficacy of Hybrid Boruta-VI across feature selection methods, while also examining the performance of the Random Forest model among different models.

The roc.test function from the pROC R package[39] is utilized to conduct DeLong's test for comparing two ROC curves and evaluating the statistical significance of the true difference in the AUC for two models. In Table 4, we compare the AUCs of various ML algorithms under two distinct FS methods. The results indicate that there is a notable difference in AUC values between the hybrid Boruta-VI method and other FS techniques across most ML algorithms. Particularly, in models such as Random Forest, C5.0, GBM, XGBOOST, and Bagged CART, where high AUC values were observed, significant differences were identified. In the case of the hybrid Boruta-VI method versus the absence of feature selection (utilizing all features), the AUC values for C5.0, GBM, XGBOOST, and Bagged CART exhibited statistically significant disparities. However, when comparing hybrid Boruta-VI with the MGD feature selection method, the AUC values were found not to be significantly different. Furthermore, when contrasting hybrid Boruta-VI with the CMIM approach, noteworthy differences in AUC values were specifically observed for C5.0 and GBM models. Moreover, for the hybrid Boruta-VI method compared to Hybrid ABL, significant disparities in AUC were identified for Bagged CART and GBM models. Lastly, when examining Hybrid Boruta-VI versus the correlated FS method, a significant difference in AUC was noted specifically for the GBM model. Notably, when applying the Random Forest model, the difference in AUC was significantly equal to 0 for Boruta-VI in comparison to various other FS methods, except for CMIM.

Table 3 and Fig. 6 present comparative performance metrics indicating that the Random Forest algorithm consistently achieved the highest accuracy across all FS methods.

To rigorously assess the significance of these results from a statistical perspective, we conducted hypothesis tests to compare the performance of Random Forest with other machine learning algorithms using the Hybrid Boruta-VI dataset (Table 5).

The chi-square test results from Table 5, provide strong evidence that the predictions of each model are not attributable to random chance, indicating a meaningful relationship between the predictors and the outcomes. Subsequently, the McNemar's Chi-squared test revealed a statistically significant difference in the paired predictions of Random Forest, Bagged CART, Decision Tree (C5.0), and Extreme Gradient Boosting (XGBoost). The notably low p-values derived from these tests further reinforce the assertion that this differentiation is not arbitrary, thereby affirming a substantial association between the prediction sets of the models under examination.

Moreover, the comparison of area under the curve (AUC) values among Random Forest, Bagged CART, and Decision Tree (C5.0) models, as indicated by their relatively large p-values, suggests that their predictive performance is not significantly distinct. Consequently, based on our analytical findings, we can confidently assert that Random Forest outperforms ten alternative models, encompassing Generalized Linear Model, LDC, Regularized Regression, k-Nearest Neighbors, Naive Bayes, SVM, CART, Stochastic Gradient Boosting, Neural Network, and XGBoost. Notably, while Random Forest demonstrates superior performance over the majority

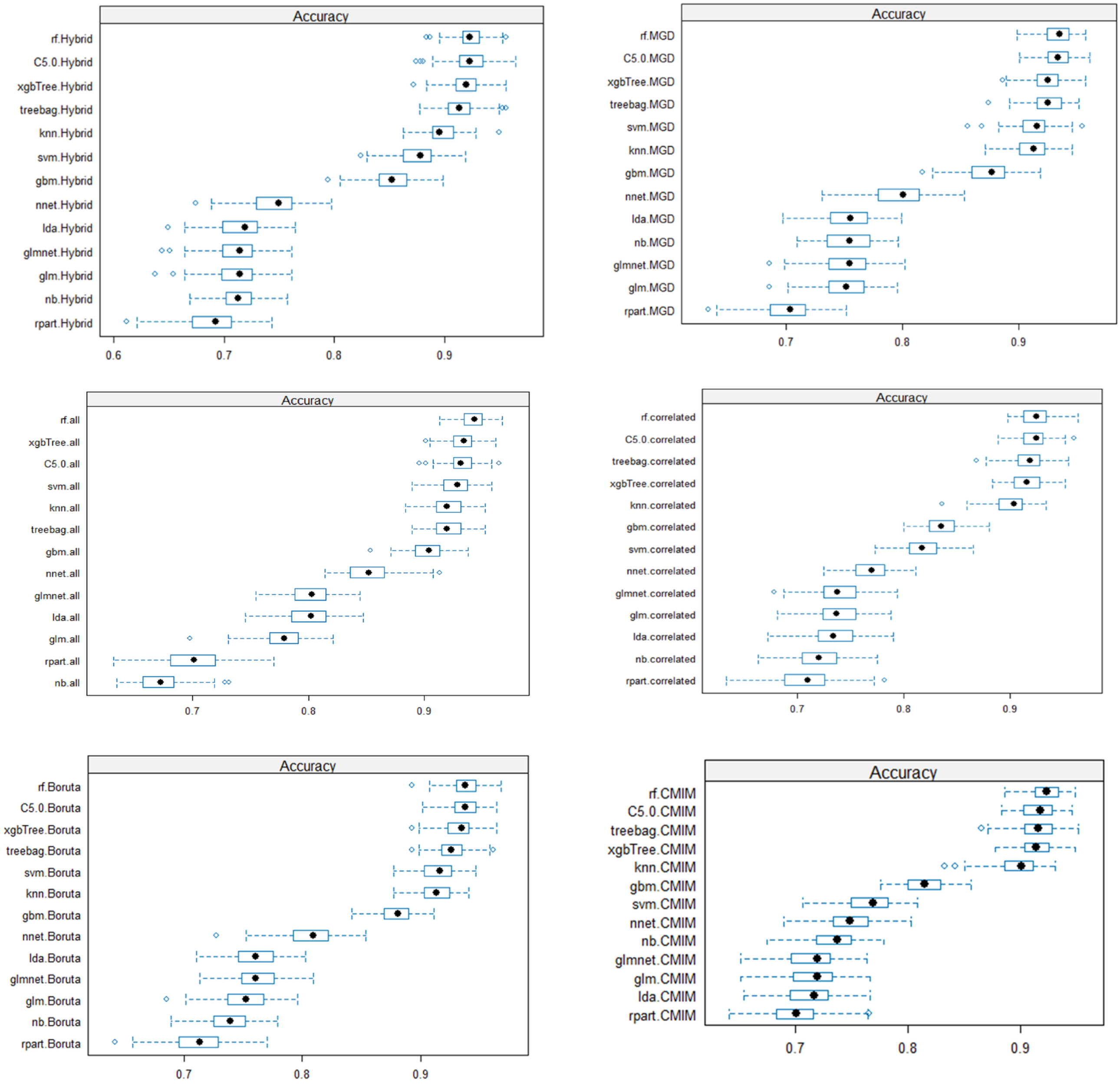

Figure 3. Comparison of performance of machine learning algorithms for different feature selection methods using 10 repeated tenfold cross validation.

of models considered, it does not exhibit a statistically significant performance advantage relative to the Bagged CART and C5.0 models.

Furthermore, the most important predictors identified by the Random Forest model on the Boruta-VI dataset include age, O2sat, UREA, ALBUMIN, CR, and LDH (Fig. 8, Supplementary Fig. S3).

Discussion

Feature Selection aims to identify a small set of features that demonstrate high classification performance[40]. This study employed both parametric and non-parametric learners in FS methods to predict COVID-19 patient mortality using hospital data.

Three types of FS methods were implemented: (i) filter methods, including CMIM and Correlation matrix; (ii) random forest embedded method; and (iii) hybrid methods, such as ABL (a combination of ANOVA, Backward selection, and Lasso methods) and Boruta-VI (a fusion of Boruta and combination of Variable Importance of multiple classification methods).

Each method revealed distinct insights into the dataset:

- CMIM identified six key features, while the correlation filter method uncovered significant correlations for eight numerical and strong Phi correlation coefficients for three categorical features.

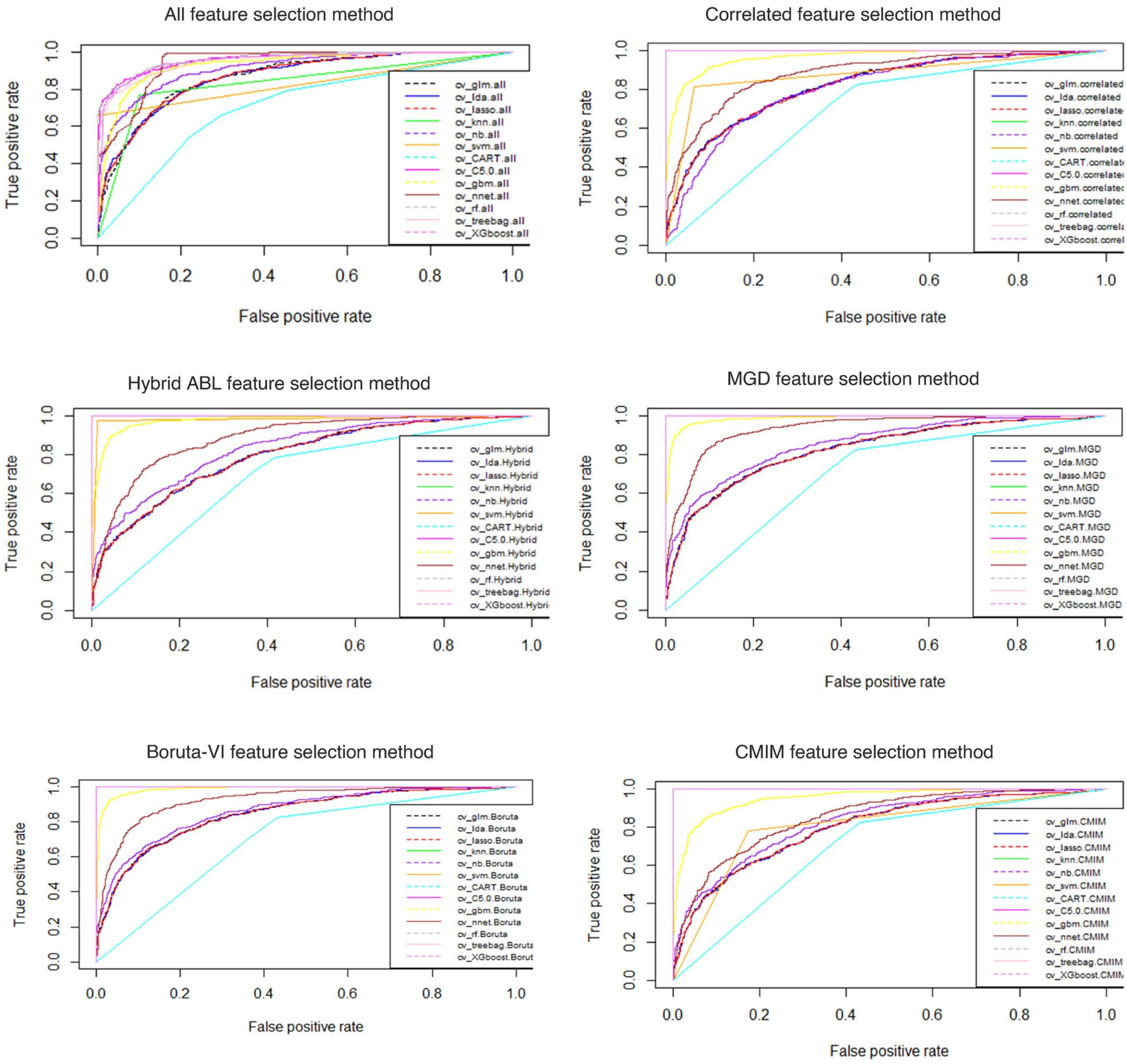

Figure 4. ROC curve for cross-validated models on the different feature selection method's data.

- The Gini impurity embedded method selected twenty important features.
- The Hybrid ABL method eliminated 15 predicting features, while the Boruta-VI hybrid method identified 20 features based on importance scores across eleven prediction models.

Comparative analysis revealed that the Hybrid Boruta-VI model outperformed other models in predicting death, showcasing their superior performance.

The results of the analysis investigating the relationship between the occurrence of Death and predictor variables in the Boruta-VI dataset present some interesting insights.

In the multivariable binary logistic regression analysis, variables such as Platelets.FFP.Injection, age, Dialysis, Decreased Consciousness, Ferritin, NEUT, P, and LACTATE exhibited the highest odds ratios (OR) in relation to Death. However, CR and UREA were not found to be significantly associated with Death at the 5% significance level in this analysis (Table 2). This finding suggests that the impact of CR and UREA on predicting Death may not be as strong in this particular model.

On the other hand, age, O2sat, ALBUMIN, UREA, Decreased Consciousness, and NEUT were highlighted as variables with high mean variable importance scores across the 11 prediction models used in the Boruta-VI hybrid feature selection method (Supplementary Table S4). These variables are considered important in predicting outcomes in the dataset across various models.

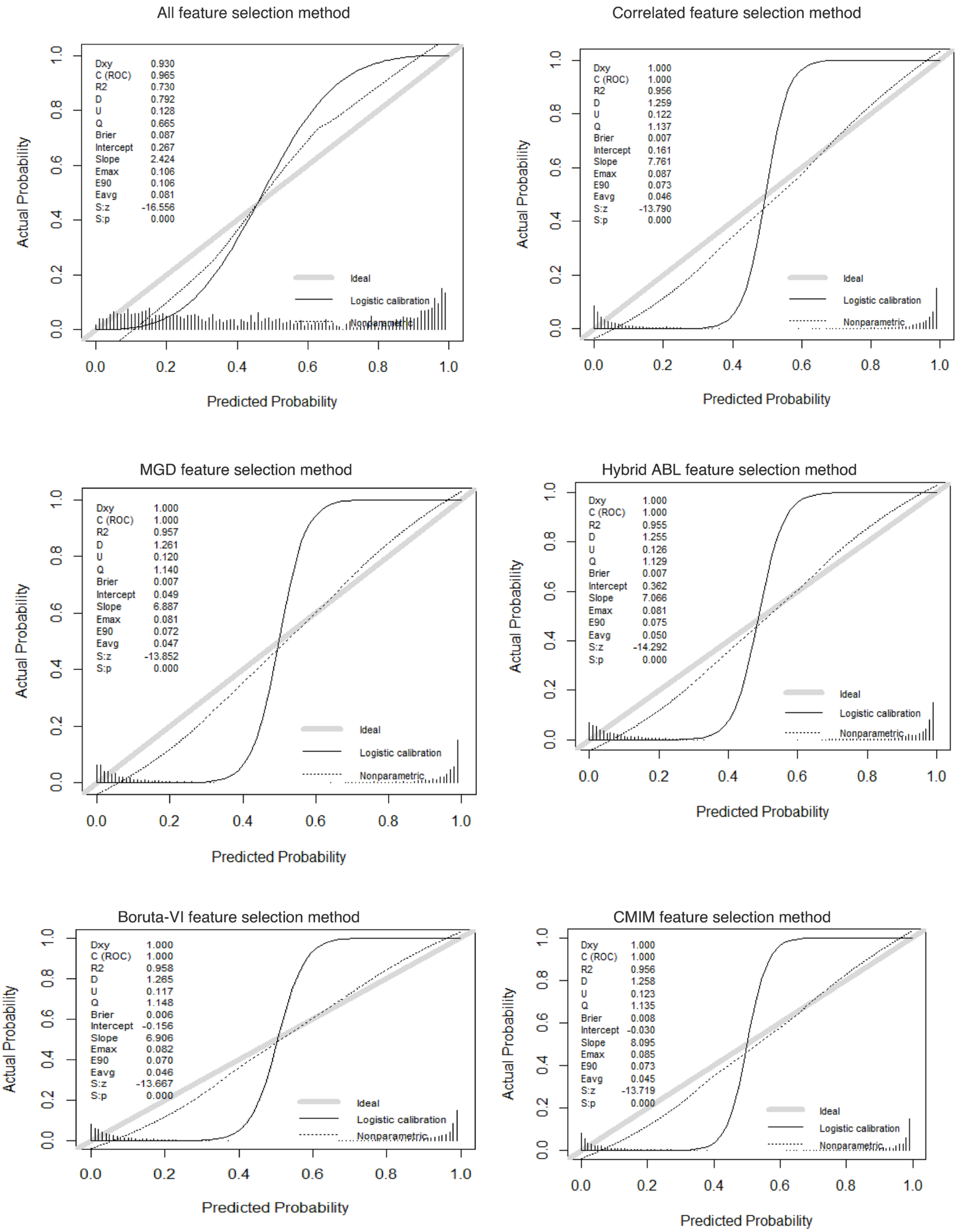

Figure 5. The performance of Random Forest model in different FS methods using repeated 10 repeated tenfold Cross Validation dataset.

Interestingly, when considering the most important predictors identified by the Random Forest model in the Boruta-VI dataset (Fig. 8), we observe a different set of key predictors. The Random Forest model places emphasis on age, O2sat, UREA, ALBUMIN, CR, and LDH variables as having high significant predictive power in determining the occurrence of Death.

Overall, these results suggest that while different models may prioritize certain variables over others in predicting Death, it is essential to consider the combined insights from various approaches.

		CMIM (6 features)	Correlation (11 features)	Hybrid ABL (15 features)	MDG Random forest (20 features)	Hybrid Boruta-VI (20 features)	Without FS (115 features)
Generalized linear model	Accuracy	0.77	0.76	0.73	0.78	0.77	0.78
	Kappa	0.38	0.39	0.36	0.43	0.47	0.47
	Sensitivity	0.70	0.68	0.69	0.70	0.72	0.76
	Specificity	0.76	0.78	0.74	0.80	0.79	0.79
	Precision	0.45	0.46	0.43	0.50	0.52	0.51
	f1-score	0.55	0.55	0.53	0.58	0.61	0.61
	AUC	0.81	0.84	0.81	0.84	0.85	0.84
Linear discriminant classifier (LDC)	Accuracy	0.75	0.75	0.74	0.78	0.80	0.79
	Kappa	0.38	0.42	0.36	0.44	0.47	0.47
	Sensitivity	0.69	0.66	0.69	0.69	0.70	0.73
	Specificity	0.76	0.78	0.75	0.81	0.80	0.81
	Precision	0.45	0.46	0.44	0.51	0.53	0.52
	f1-score	0.55	0.54	0.53	0.58	0.61	0.61
	AUC	0.81	0.83	0.81	0.84	0.85	0.84
Lasso regression	Accuracy	0.75	0.75	0.73	0.83	0.79	0.79
	Kappa	0.38	0.39	0.36	0.42	0.47	0.47
	Sensitivity	0.70	0.67	0.69	0.68	0.72	0.75
	Specificity	0.76	0.78	0.74	0.80	0.81	0.79
	Precision	0.45	0.46	0.43	0.50	0.52	0.52
	f1-score	0.55	0.55	0.53	0.58	0.61	0.61
	AUC	0.81	0.83	0.81	0.84	0.85	0.84
k-nearest neighbors	Accuracy	0.83	0.83	0.82	0.86	0.85	0.86
	Kappa	0.57	0.57	0.53	0.61	0.61	0.61
	Sensitivity	0.81	0.76	0.75	0.76	0.76	0.73
	Specificity	0.82	0.85	0.84	0.88	0.88	0.90
	Precision	0.58	0.60	0.57	0.65	0.65	0.68
	f1-score	0.68	0.67	0.65	0.70	0.70	0.70
	AUC	0.83	0.83	0.82	0.84	0.84	0.82
Naive Bayes	Accuracy	0.75	0.81	0.77	0.80	0.81	0.81
	Kappa	0.38	0.41	0.36	0.43	0.41	0.34
	Sensitivity	0.67	0.49	0.59	0.55	0.50	0.28
	Specificity	0.77	0.89	0.81	0.87	0.89	0.94
	Precision	0.45	0.57	0.49	0.55	0.57	0.71
	f1-score	0.54	0.53	0.52	0.55	0.51	0.40
	AUC	0.82	0.82	0.81	0.82	0.84	0.85
Support vector machine	Accuracy	0.77	0.82	0.77	0.83	0.84	0.88
	Kappa	0.42	0.49	0.43	0.55	0.60	0.75
	Sensitivity	0.68	0.68	0.72	0.76	0.83	0.78
	Specificity	0.76	0.85	0.78	0.85	0.85	0.83
	Precision	0.48	0.56	0.49	0.58	0.61	0.60
	f1-score	0.57	0.62	0.58	0.66	0.70	0.79
	AUC	0.72	0.77	0.84	0.82	0.86	0.79
Classification and regression trees	Accuracy	0.63	0.63	0.61	0.62	0.63	0.68
	Kappa	0.24	0.24	0.22	0.26	0.24	0.27
	Sensitivity	0.84	0.74	0.75	0.84	0.84	0.65
	Specificity	0.55	0.60	0.57	0.55	0.55	0.69
	Precision	0.34	0.34	0.33	0.34	0.34	0.37
	f1-score	0.49	0.47	0.46	0.49	0.49	0.48
	AUC	0.69	0.67	0.66	0.68	0.67	0.70
C5.0 decision tree	Accuracy	0.86	0.88	0.88	0.89	0.89	0.89
	Kappa	0.64	0.68	0.65	0.70	0.70	0.69
	Sensitivity	0.81	0.84	0.77	0.81	0.84	0.83
	Specificity	0.88	0.89	0.91	0.92	0.90	0.90
	Precision	0.66	0.68	0.70	0.72	0.70	0.71
	f1-score	0.72	0.75	0.73	0.76	0.77	0.76
	AUC	0.92	0.94	0.95	0.95	0.95	0.81
Continued							

		CMIM (6 features)	Correlation (11 features)	Hybrid ABL (15 features)	MDG Random forest (20 features)	Hybrid Boruta-VI (20 features)	Without FS (115 features)
Random Forest	Accuracy	0.89	0.88	0.88	0.88	0.88	0.90
	Kappa	0.64	0.68	0.67	0.68	0.69	0.71
	Sensitivity	0.82	0.85	0.80	0.81	0.83	0.79
	Specificity	0.88	0.89	0.91	0.91	0.91	0.93
	Precision	0.66	0.68	0.70	0.71	0.71	0.76
	f1-score	0.73	0.76	0.75	0.76	0.76	0.78
	AUC	0.93	0.94	0.94	0.95	0.95	0.96
Neural network	Accuracy	0.74	0.76	0.73	0.77	0.78	0.78
	Kappa	0.34	0.42	0.35	0.45	0.49	0.49
	Sensitivity	0.71	0.71	0.65	0.74	0.76	0.87
	Specificity	0.74	0.76	0.75	0.78	0.81	0.75
	Precision	0.43	0.47	0.43	0.49	0.53	0.50
	f1-score	0.54	0.57	0.53	0.59	0.62	0.63
	AUC	0.81	0.82	0.82	0.82	0.85	0.79
Stochastic gradient boosting	Accuracy	0.81	0.81	0.81	0.84	0.85	0.87
	Kappa	0.51	0.52	0.51	0.59	0.61	0.66
	Sensitivity	0.76	0.78	0.73	0.81	0.84	0.84
	Specificity	0.82	0.82	0.83	0.85	0.85	0.88
	Precision	0.54	0.56	0.56	0.61	0.61	0.67
	f1-score	0.63	0.64	0.63	0.69	0.71	0.75
	AUC	0.87	0.88	0.88	0.89	0.90	0.84
Bagged CART	Accuracy	0.85	0.86	0.85	0.87	0.88	0.88
	Kappa	0.62	0.64	0.60	0.66	0.68	0.68
	Sensitivity	0.82	0.85	0.79	0.83	0.86	0.84
	Specificity	0.86	0.86	0.86	0.88	0. 89	0.89
	Precision	0.63	0.64	0.61	0.66	0.69	0.68
	f1-score	0.71	0.73	0.70	0.74	0.76	0.76
	AUC	0.93	0.94	0.94	0.94	0.95	0.79
Extreme gradient boosting	Accuracy	0.86	0.87	0.85	0.87	0.88	0.88
	Kappa	0.63	0.66	0.61	0.67	0.67	0.67
	Sensitivity	0.79	0.84	0.76	0.82	0.83	0.82
	Specificity	0.88	0.88	0.88	0.89	0.89	0.89
	Precision	0.66	0.67	0.64	0.68	0.68	0.69
	f1-score	0.72	0.74	0.69	0.75	0.75	0.75
	AUC	0.91	0.93	0.93	0.93	0.93	0.83

Table 3. Performance comparison between six FS methods for different classifiers.

Multiple studies have consistently demonstrated that age is a critical risk factor for severe outcomes and increased mortality rates among COVID-19 patients. Older individuals are more likely to experience complications and have a higher mortality rate due to age-related declines in immune function and increased prevalence of comorbidities. For example, a study by Xu et al.[41] reported that the median age of non-survivors was significantly higher than survivors, highlighting the association between age and COVID-19 mortality.

However, it is important to consider how age interacts with other variables in the model and whether the impact of age on outcomes is consistent across different demographic groups. Failure to account for potential age-related biases can result in inequalities in healthcare resource allocation and treatment decisions.

Hypoxemia, indicated by reduced oxygen saturation levels, is a hallmark of severe COVID-19 and is associated with respiratory distress and the need for intensive care. Several studies, including the one by Wu et al.[42], have highlighted the prognostic significance of O2sat as a predictor of disease severity and mortality in COVID-19 patients.

Low serum albumin levels have been consistently associated with worse clinical outcomes in COVID-19 patients. Albumin serves as a marker of nutritional status and overall health, and hypoalbuminemia has been linked to increased disease severity and mortality. Studies, such as the one by Alirezaei et al.[43], have demonstrated the prognostic value of albumin in predicting adverse outcomes in COVID-19 patients.

Elevated neutrophil counts have been linked to severe COVID-19 cases and poor outcomes. Neutrophilia is often observed in patients with cytokine storms, a hyper inflammatory response associated with severe COVID-19. Studies, such as the one by Liu et al.[44], have demonstrated a correlation between increased neutrophil levels and disease severity, suggesting that NEUT is a valuable predictor of adverse outcomes.

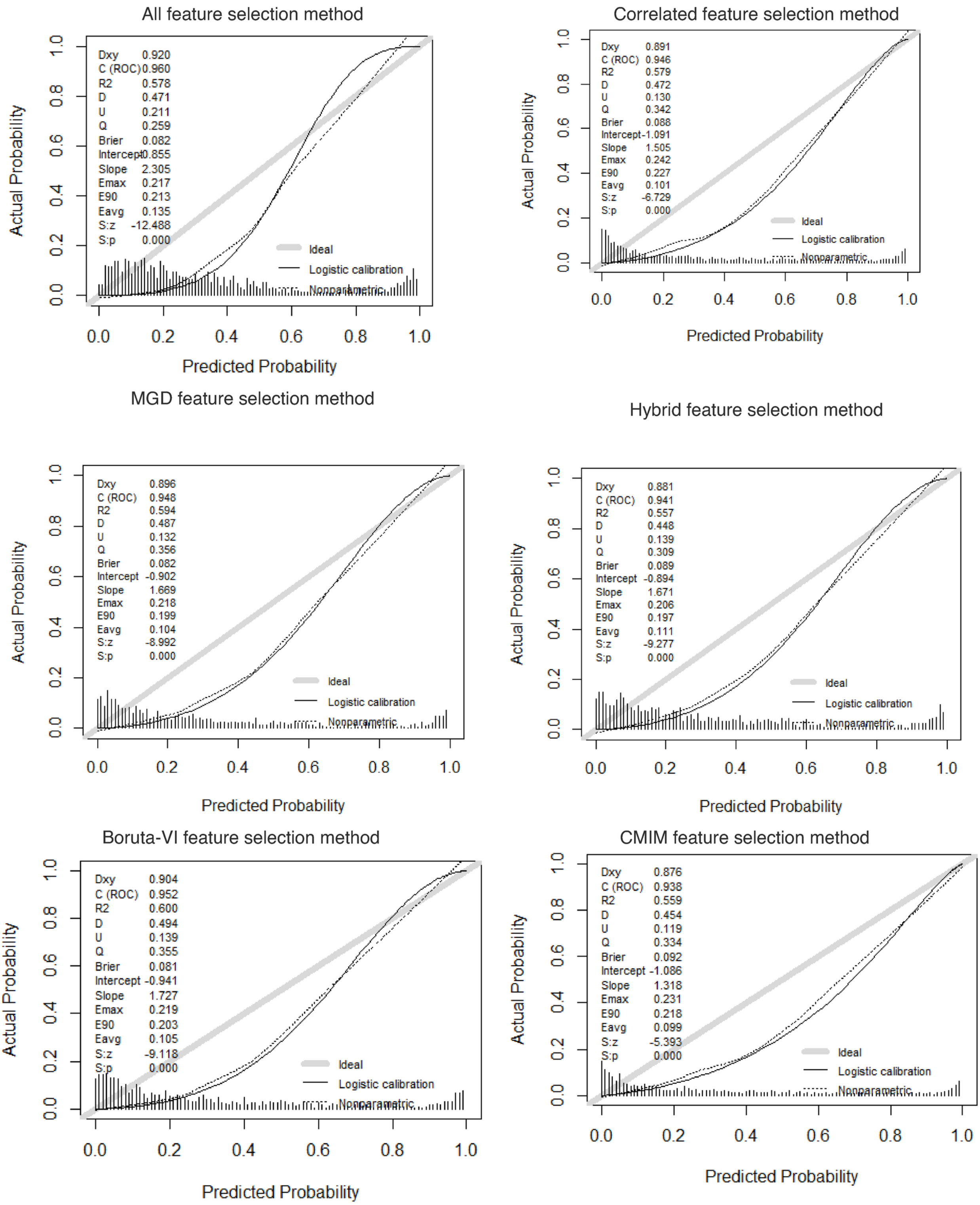

Figure 6. The performance of Random Forest model in different feature selection methods using test data (external validation).

The correlation matrix on Hybrid Boruta-VI dataset emphasized significant relationships between various parameters, such as UREA and CR, as well as UREA and BUN, shedding light on their importance in assessing kidney function. Additionally, connections between UREA levels and phosphate levels, as well as heart function, were discerned through positive correlations with P and PROBNP.

Elevated levels of urea and BUN have been identified as predictors of poor prognosis in COVID-19 patients. Abnormal kidney function, indicated by elevated urea and BUN, is associated with an increased risk of severe

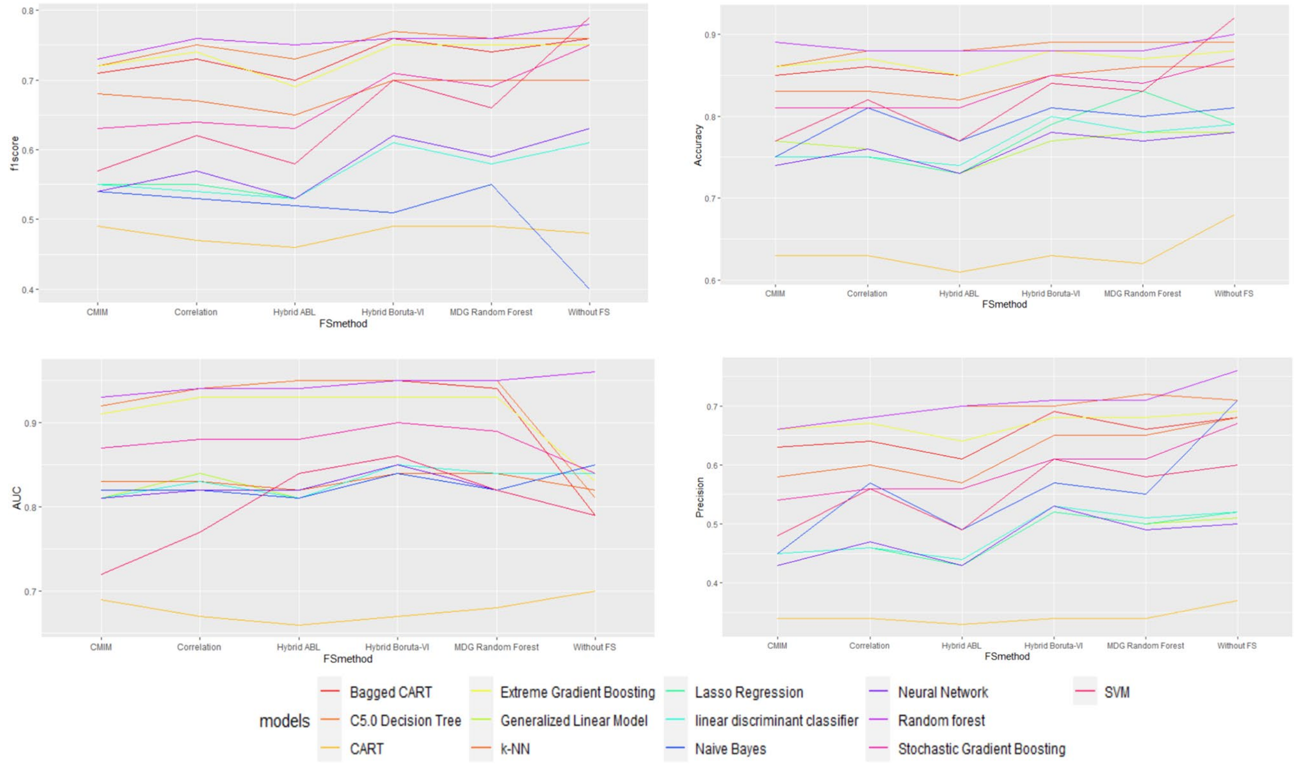

Figure 7. Comparing the F1-score, accuracy, AUC and precision of various machine learning models across different feature selection methods.

	Hybrid Boruta-VI, all		Hybrid Boruta-VI, correlated		Hybrid Boruta-VI, MGD		Hybrid Boruta-VI, Hybrid ABL Boruta-VI		Hybrid Boruta-VI, CMIM	
	Z (95 percent confidence interval)	p-value	Z (95 percent confidence interval)	p-value	Z (95 percent confidence interval)	p-value	Z (95 percent confidence interval)	p-value	Z (95 percent confidence interval)	p-value
Generalized linear model	0.27 [− 0.01, 0.02]	0.78	1.15 [− 0.01, 0.02]	0.25	0	1	4.37 [0.02, 0.06]	1.228e−05	1.84 [− 0.93, 4.61]	0.06
Linear discriminant classifier (LDC)	0.09 [− 0.01, 0.02]	0.93	3.05 [0.01, 0.03]	0.00	2.76 [0.00, 0.03]	0.00	4.23 [0.02, 0.06]	2.333e−05	2.41 [− 0.36, 5.18]	0.02
Regularized regression	− 0.20 [− 0.02, 0.01]	0.84	3.01 [0.01, 0.03]	0.00	2.67 [0.00, 0.03]	0.007	4.18 [0.02, 0.06]	2.91e−05	2.34 [− 0.42, 5.12]	0.02
k – nearest neighbors	1.03 [− 1.74, 3.80]	0.30	0.62 [− 0.35, 1.58]	0.53	− 0.60 [− 0.02, 0.01]	0.55	1.10 [− 0.01, 0.04]	0.26	16.17 [13.4, 18.94]	< 2.2e−16
Naive Bayes	− 0.62 [− 3.39, 2.15]	0.53	2.24 [0.00 ,0.04]	0.02	2.62 [0.00, 0.02]	0.01	2.97 [0.01, 0.05]	0.00	0.84 [− 1.93, 3.61]	0.40
Support vector machine	3.78 [0.03, 0.10]	0.00	5.99 [0.06, 0.11]	1.976e−09	0.57 [− 0.01, 0.03]	0.57	2.88 [0.01, 0.06]	0.00	7.60 [0.09, 0.15]	2.942e−14
Classification and regression trees (CART)	− 1.06 [− 0.04, 0.01]	0.29	0	1	0	1	1.14 [− 0.00, 0.017]	0.25	− 0.41 [− 3.18, 2.36]	0.68
Stochastic gradient boosting (GBM)	6.39 [0.041, 0.08]	1.613e−10	3.32 [0.00, 0.04]	0.00	1.43 [− 0.00, 0.01]	0.15	2.86 [0.01, 0.04]	0.00	2.12 [− 0.65, 4.89]	0.03
Neural network	3.41 [0.63, 6.20]	0.00	2.84 [0.01, 0.06]	0.00	3.10 [0.01, 0.06]	0.00	2.65 [0.01, 0.06]	0.1	2.64 [− 0.50, 5.05]	0.02
Decision tree (C5.0)	11.79 [0.12, 0.16]	2.2e−16	1.81 [− 0.000, 0.02]	0.07	0.94 [− 0.00, 0.01]	0.34	0.71 [− 0.01, 0.01]	0.48	2.64 [− 0.13, 5.41]	0.01
Random Forest	− 0.58 [− 3.36, 2.20]	0.56	1.52 [− 0.00, 0.01]	0.13	1.11 [− 0.00, 0.01]	0.26	0.19 [− 0.01, 0.01]	0.85	22.07 [19.30, 24.84]	< 2.2e−16
Bagged CART	12.83 [0.13, 0.18]	< 2.2e−16	1.42 [− 0.00, 0.02]	0.15	1.81 [− 0.00, 0.01]	0.071	2.24 [0.00, 0.02]	0.02	1.54 [− 1.22, 4.38]	0.12
Extreme gradient boosting (XGBoost)	9.20 [0.08, 0.13]	< 2.2e−16	0.82 [− 0.01, 0.020]	0.41	− 0.09 [− 0.01, 0.01]	0.93	0.28 [− 0.01, 0.01]	0.78	1.64 [− 1.13, 4.42]	0.10

Table 4. Comparing AUCs of machine learning models for two feature selections method with DeLong's test (Roc.test).

	chisq.test		mcnemar.test		roc.test	
	X-squared	p-value	McNemar's chi-squared	p-value	Z (95 percent confidence interval)	p-value
RF, generalized linear model	604.66	<2.2e−16	23.902	1.013e−06	−7.83 (−10.61, −5.06)	7.29e−15
RF, linear discriminant classifier (LDC)	613.78	<2.2e−16	11.695	0.00	−7.55 (−10.33, −4.80)	6.102e−14
RF, regularized regression	608.7	<2.2e−16	24.71	6.661e−07	−7.57 (−10.34, −4.80)	6.102e−14
RF, k-nearest neighbors	332.71	<2.2e−16	97.352	<2.2e−16	−12.804 (−0.17, −0.12)	<2.2e−16
RF, Naive Bayes	544.87	<2.2e−16	29.952	4.428e−08	−11.29 (−0.15, −0.10)	<2.2e−16
RF, support vector machine	602.4	<2.2e−16	40.416	2.053e−10	−10.7 (−0.13, −0.09)	<2.2e−16
RF, classification and regression trees (CART)	209.69	<2.2e−16	246.24	<2.2e−16	−19.85 (−0.32, −0.26)	<2.2e−16
RF, stochastic gradient boosting	819.04	<2.2e−16	31.89	1.632e−08	−13.09 (−0.18, −0.14)	<2.2e−16
RF, neural network	425.80	<2.2e−16	38.162	6.511e−10	−13.09 (−0.18, −0.14)	<2.2e−16
RF, decision tree (C5.0)	1011	<2.2e−16	0.73	0.39	−1.07 (−0.01, 0.00)	0.2852
Random forest, bagged CART	1107.7	<2.2e−16	10.09	0.00	−2.46 (−0.01, −0.00)	0.01375
Random forest, extreme gradient boosting (XGBoost)	1031.1	<2.2e−16	2.3614	0.1244	−4.83 (−0.03, −0.01)	1.316e−06

Table 5. Statistical tests for pairwise performance comparison between different models and Random Forest on the hybrid Boruta-VI dataset.

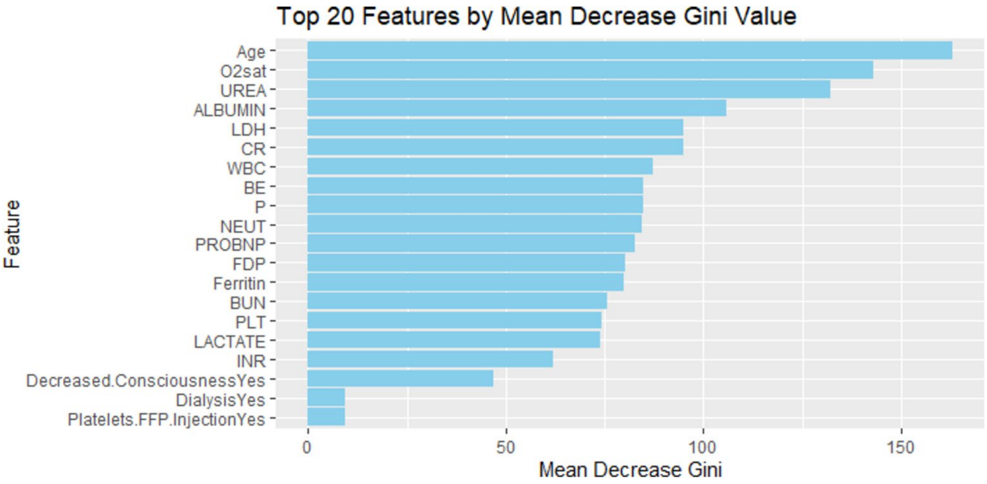

Figure 8. Feature importance of the Random Forest model on Boruta-VI dataset.

complications and mortality. Research, such as the study by Liu et al.[45], have shown that renal dysfunction, as indicated by elevated urea and BUN, is a significant risk factor for adverse outcomes in COVID-19.

Comorbidities, such as diabetes, hypertension, or cardiovascular disease, are common risk factors for severe COVID-19 outcomes. Prediction models that focus solely on comorbidities as predictors may disproportionately impact individuals with pre-existing health conditions. Considering the prevalence of comorbidities across different demographic groups is essential to evaluate the fairness of the model. Ensuring that the predictive power of comorbidities is consistent and unbiased across diverse populations can help prevent disparities in healthcare outcomes.

The evaluation of model performance across different FS methods using various machine learning algorithms favored the Random Forest model consistently. While the Hybrid Boruta-VI method displayed competitive performance, it did not consistently outperform other methods across all metrics. Comparative analyses and hypothesis tests further reinforced the effectiveness of the Random Forest model in predicting outcomes in comparison to various other machine learning algorithms.

The RF model's ability to handle high-dimensional data, handle non-linear relationships, handle feature interactions, and mitigate the impact of noisy or irrelevant features may contribute to the superior performance of it compared to other models.

The accurate prediction of COVID-19 mortality using machine learning models has significant clinical utility and implications for patient care and decision-making. These predictive models can assist healthcare professionals in making informed decisions about resource allocation, treatment strategies, and patient management. Syed et al.[46] used a hybrid approach to identify blood biomarkers predicting COVID-19 mortality. They employed mRMR, t-test, and whale optimization algorithm (WOA) for FS and trained ML algorithms, finding that the RF model accurately predicts mortality with an accuracy of 0.96, F1 score of 0.96, and AUC value of 0.98 on independent test data.

Brinati et al.[47] developed a model for predicting COVID-19 using various ML approaches including DT, KNN, RF, LR, NB, SVM, and trees weighting RF (TWRF). The RF classifier demonstrated the best performance, achieving an AUC of 84%, accuracy of 82%, sensitivity of 92%, and specificity of 65%.

Liang et al.[48] developed a risk score model to predict critical illness. It identified 19 important predictors using LASSO regression and achieved an AUC of 0.88.

Amini et al.[49] utilized five FS methods in prediction of COVID-19 mortality, including Forward FS, minimum Redundancy Maximum Relevance, Relief, Linear Discriminant Analysis, and Neighborhood Component Analysis. Their findings indicated that the RF classifier, combined with Forward FS, achieved the highest classification accuracy of 92.08 ± 2.56.

Most of these studies demonstrated superior performance using the RF classifier combined with hybrid methods. However, it is important to note that the differences in performance could be attributed to various factors, including variations in data characteristics. While our study utilized a balanced dataset, many of the previous studies employed imbalanced datasets, which may have influenced their results. Imbalanced datasets can introduce biases and affect the performance metrics of predictive models, potentially leading to inflated accuracy or other performance measures.

It is worth noting that using a balanced dataset in our study allows for a more accurate assessment of the model's performance in predicting COVID-19 mortality. However, it is crucial to acknowledge that real-world datasets often exhibit class imbalance, where the number of COVID-19 mortality cases is significantly lower than the number of non-mortality cases. Therefore, future research should explore the performance of predictive models on imbalanced datasets and investigate techniques for handling class imbalance effectively.

The study has limitations and strengths in its methodology and findings related to the use of machine learning models for predicting outcomes in COVID-19 patients. Some key limitations include the potential lack of generalizability to broader populations, the retrospective nature of the study, and the single-center focus on patients from Tehran, Iran. The need for external validation on larger and multi-center databases is highlighted as important for future investigations.

On the other hand, the study also leveraged strengths such as utilizing COVID-19 data from multiple hospitals in Tehran, employing robust scaling methods, having a balanced dataset, and utilizing a variety of feature selection methods and machine learning algorithms. Measures were taken to mitigate overfitting risks through repeated cross-validation and maintaining an adequate sample size for predictor variables.

Moving forward, it will be essential for future studies to address the identified limitations, especially through external validations on diverse populations, prospective cohort evaluations, and assessing the real-world impact of predictive models on medical decision-making and patient outcomes. This iterative process of validation and improvement is crucial for the successful implementation of machine learning models in clinical settings.

In conclusion, the combination of different FS methods can benefit research and clinical applications in predicting COVID-19 outcomes by providing a more comprehensive understanding of important predictors, overcoming method limitations, and reducing the impact of bias or confounding factors. The predictive models developed in the study have significant clinical utility and implications for patient care and decision-making. Healthcare professionals can use these models to make informed decisions about resource allocation, treatment strategies, and patient management. However, considerations should be made regarding how easily the model can be integrated into existing care practices, the level of expertise required for users to interpret and utilize the model, and any potential ethical implications of using predictive models in clinical settings.

Data availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Received: 3 January 2024; Accepted: 1 August 2024
Published online: 11 August 2024

References

1. Mishra, S. & Pradhan, R. K. Analyzing the impact of feature correlation on classification acuracy of machine learning model. In *2023 International Conference on Artificial Intelligence and Smart Communication (AISC)* (2023).
2. Chandrashekar, G. & Sahin, F. A survey on feature selection methods. *Comput. Electr. Eng.* **40**(1), 16–28 (2014).
3. Venkatesh, B. & Anuradha, J. A review of feature selection and its methods. *Cybern. Inf. Technol.* **19**(1), 3–26 (2019).
4. Pudjihartono, N., Fadason, T., Kempa-Liehr, A. W. & O'Sullivan, J. M. A review of feature selection methods for machine learning-based disease risk prediction. *Front. Bioinform.* **2**, 927312 (2022).
5. Uppu, S., Krishna, A. & Gopalan, R. P. A review on methods for detecting SNP interactions in high-dimensional genomic data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **15**(2), 599–612 (2016).
6. Ali, R. H. & Abdulsalam, W. H. The prediction of COVID 19 disease using feature selection techniques. *J. Phys. Conf. Ser.* **1879**, 1 (2021).
7. Pourhomayoun, M. & Shakibi, M. Predicting mortality risk in patients with COVID-19 using machine learning to help medical decision-making. *Smart Health* **20**, 100178 (2021).
8. Varzaneh, Z. A., Orooji, A., Erfannia, L. & Shanbehzadeh, M. A new COVID-19 intubation prediction strategy using an intelligent feature selection and K-NN method. *Inform. Med. Unlocked* **28**, 100825 (2022).
9. Hayet-Otero, M. *et al.* Extracting relevant predictive variables for COVID-19 severity prognosis: An exhaustive comparison of feature selection techniques. *PLoS ONE* **18**(4), e0284150 (2023).
10. Chamseddine, E., Mansouri, N., Soui, M. & Abed, M. Handling class imbalance in COVID-19 chest X-ray images classification: Using SMOTE and weighted loss. *Appl. Soft Comput.* **129**, 109588 (2022).

11. Javidi, M., Abbaasi, S., Naybandi Atashi, S. & Jampour, M. COVID-19 early detection for imbalanced or low number of data using a regularized cost-sensitive CapsNet. *Sci. Rep.* **11**(1), 18478 (2021).
12. Hatamabadi, H. *et al.* Epidemiology of COVID-19 in Tehran, Iran: A cohort study of clinical profile, risk factors, and outcomes. *BioMed Res. Int.* **2022**, 2350063 (2022).
13. Sharma, V. A study on data scaling methods for machine learning. *Int. J. Glob. Acad. Sci. Res.* **1**(1), 23–33 (2022).
14. Zali, A. *et al.* Baseline characteristics and associated factors of mortality in COVID-19 patients: An analysis of 16000 cases in Tehran, Iran. *Arch. Acad. Emerg. Med.* **8**(1), e70 (2020).
15. Ogundimu, E. O., Altman, D. G. & Collins, G. S. Adequate sample size for developing prediction models is not simply related to events per variable. *J. Clin. Epidemiol.* **76**, 175–182 (2016).
16. Alin, A. Multicollinearity. *Wiley interdiscip. Rev. Comput. Stat.* **2**(3), 370–374 (2010).
17. Daoud, J. I. Multicollinearity and regression analysis. *J. Phys. Conf. Ser.* **949**, 1 (2017).
18. Vidal-Naquet, M. & Ullman, S. (eds) *Object Recognition with Informative Features and Linear Classification* (ICCV, 2003).
19. Fleuret, F. Fast binary feature selection with conditional mutual information. *J. Mach. Learn. Res.* **5**, 9 (2004).
20. Bommert, A., Welchowski, T., Schmid, M. & Rahnenführer, J. Benchmark of filter methods for feature selection in high-dimensional gene expression survival data. *Brief. Bioinform.* **23**(1), 354 (2022).
21. Schratz, P. L. M. & Bischl, B. *mlr3filters: Filter Based Feature Selection for 'mlr3'* (2020).
22. Bommert, A., Sun, X., Bischl, B., Rahnenführer, J. & Lang, M. Benchmark for filter methods for feature selection in high-dimensional classification data. *Comput. Stat. Data Anal.* **143**, 106839 (2020).
23. Menze, B. H. *et al.* A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinform.* **10**(1), 213 (2009).
24. Nembrini, S., König, I. R. & Wright, M. N. The revival of the Gini importance? *Bioinformatics* **34**(21), 3711–3718 (2018).
25. Han, H., Guo, X. & Yu, H. Variable selection using mean decrease accuracy and mean decrease Gini based on random forest. In *2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS)* (IEEE, 2016).
26. Sheskin, D. J. *Handbook of Parametric and Nonparametric Statistical Procedures* (CRC Press, 2020).
27. Moorthy, U. & Gandhi, U. D. A novel optimal feature selection technique for medical data classification using ANOVA based whale optimization. *J. Amb. Intell. Hum. Comput.* **12**, 3527–3538 (2021).
28. Ladha, L. *et al.* Feature selection methods and algorithms. *Int. J. Comput. Sci. Eng.* **1**, 1 (2022).
29. Okser, S. *et al.* Regularized machine learning in the genetic prediction of complex traits. *PLoS Genet.* **10**(11), e1004754 (2014).
30. Kursa, M. B. & Rudnicki, W. R. Feature selection with the Boruta package. *J. Stat. Softw.* **36**, 1–13 (2010).
31. Bottino, F. *et al.* COVID mortality prediction with machine learning methods: A systematic review and critical appraisal. *J. Pers. Med.* **11**, 9 (2021).
32. Berrar, D. *Cross-Validation* (2019).
33. Kuhn, M. Building predictive models in R using the caret package. *J. Stat. Softw.* **28**, 1–26 (2008).
34. Kuhn, M. *Variable Selection Using the Caret Package.* http://cran.cermin.lipi.go.id/web/packages/caret/vignettes/caretSelection.pdf (2012).
35. Krawczyk, B. Learning from imbalanced data: Open challenges and future directions. *Prog. Artif. Intell.* **5**(4), 221–232 (2016).
36. Lunardon, N., Menardi, G. & Torelli, N. ROSE: A package for binary imbalanced learning. *R J.* **6**(1), 79 (2014).
37. Jaccard, P. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bull. Soc. Vaudoise Sci. Nat.* **37**, 547–579 (1901).
38. Wei, T. *et al.* Package 'corrplot'. *Statistician* **56**(316), e24 (2017).
39. Robin, X. *et al. Package 'pROC'. Package "pROC"* (2021).
40. Tang, J., Alelyani, S. & Liu, H. Feature selection for classification: A review. In *Data Classification: Algorithms and Applications* 37 (2014).
41. Xu, W. *et al.* Risk factors analysis of COVID-19 patients with ARDS and prediction based on machine learning. *Sci. Rep.* **11**(1), 2933 (2021).
42. Wu, C. *et al.* Risk factors associated with acute respiratory distress syndrome and death in patients with coronavirus disease 2019 pneumonia in Wuhan, China. *JAMA Intern. Med.* **180**(7), 934–943 (2020).
43. Alirezaei, T. *et al.* The role of blood urea nitrogen to serum albumin ratio in the prediction of severity and 30-day mortality in patients with COVID-19. *Health Sci. Rep.* **5**(3), e606 (2022).
44. Liu, Y. *et al.* Neutrophil-to-lymphocyte ratio as an independent risk factor for mortality in hospitalized patients with COVID-19. *J. Infect.* **81**(1), e6–e12 (2020).
45. Liu, Y.-F. *et al.* The chronic kidney disease and acute kidney injury involvement in COVID-19 pandemic: A systematic review and meta-analysis. *PLoS ONE* **16**(1), e0244779 (2021).
46. Syed, A. H., Khan, T. & Alromema, N. A hybrid feature selection approach to screen a novel set of blood biomarkers for early COVID-19 mortality prediction. *Diagnostics* **12**, 7 (2022).
47. Brinati, D. *et al.* Detection of COVID-19 infection from routine blood exams with machine learning: A feasibility study. *J. Med. Syst.* **44**, 1–12 (2020).
48. Liang, W. *et al.* Development and validation of a clinical risk score to predict the occurrence of critical illness in hospitalized patients with COVID-19. *JAMA Intern. Med.* **180**(8), 1081–1089 (2020).
49. Amini, N. *et al.* Automated prediction of COVID-19 mortality outcome using clinical and laboratory data based on hierarchical feature selection and random forest classifier. *Comput. Methods Biomech. Biomed. Eng.* **26**(2), 160–173 (2023).

Acknowledgements

The authors thank the anonymous reviewers for their valuable suggestions.

Author contributions

Conceptualization: FM, SH, and KK. Data acquiring and analyzing: FM, SH, and MP. Formal analysis: FM, SH, and KK . Methodology: FM, MP, SH,KK,MZ. Project administration: FM, SH, and KK. Writing—original draft: FM, MP. Writing—review & editing: FM, MP, SH,KK,MZ. All authors reviewed the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-69209-6.

Correspondence and requests for materials should be addressed to F.M. or K.K.

Reprints and permissions information is available at www.nature.com/reprints.