A New stacking based ensemble Predictive Model to Predict the Mortality after heart surgery on a highly imbalanced dataset

ARMAN GHAVIDEL^a, ROUZBEH GHOSI^b, AND ALIREZA ATASHI^{c,d}

^a School of Industrial Engineering, Iran University of Science and Technology, Tehran 16846-13114, Iran

^b School of Industrial Engineering, Iran University of Science and Technology, Tehran 16846-13114, Iran

^c e-Health Department, Virtual School, Tehran University of Medical Sciences, Tehran 19617-33114, Iran

^d Cancer Informatics Research Group, Clinical Research Department, Breast Cancer Research Center, Motamed Cancer Institute, ACECR, Tehran 14 - 4364, Iran

Corresponding author: Rouzbeh Ghosi (ghosi@iust.ac.ir)

Abstract: Nowadays, according to spectacular improvement in health care and biomedical level, a tremendous amount of data is recorded by hospitals. In addition, the most effective approach to reduce disease mortality is to diagnose it as soon as possible. As a result, data mining by applying machine learning in the field of diseases provides good opportunities to examine the hidden patterns of this collection. An exact forecast of the mortality after heart surgery will cause Successful medical treatment and fewer costs. This research wants to recommend a new stacking predictive model after utilizing the random forest feature importance method to foresee the mortality after heart surgery on a highly unbalanced dataset by using the most practical features. To solve the unbalanced data problem, a combination of the SVM-SMOTE over-sampling algorithm and the Edited-Nearest-Neighbor under-sampling algorithm is used. This research compares the introduced model with some other machine learning classifiers to ensure efficiency through shuffle hold-out and 10-fold cross-validation strategies. In order to validate the performance of the implemented machine learning methods in this research, both shuffle hold-out, and 10-fold cross-validation results indicated that our model had the highest efficiency compared to the other models. Furthermore, the Friedman statistical test is applied to survey the differences between models. The result demonstrates that the introduced stacking model reached the most accurate predicting performance after Logistic Regression.

Keywords: Classification, stacking ensemble method, heart surgery, unbalanced data problem, hybrid predictive model, machine learning in healthcare, resampling method, Edited-Nearest-Neighbor, nonparametric test.

1. INTRODUCTION

Today, in medical knowledge, collecting an immense amount of data about different diseases is very important. Because it can lead to saving many lives and improving quality of living by discovering relations and hidden patterns among disease's features[1]. The healthcare industry is permanently generating large amounts of data, and there is a wide gap between data collection and data interpretation. Machine learning is a helpful tool that can benefit the

industry from in-depth data analysis to the development of medical research and scientific decisions in the field of diagnosis and treatment. Diagnosis and determination of appropriate treatment for patients are of great importance in medical science. In addition to wasting time and money, choosing the wrong treatment for the patient, can also have detrimental effects and, in some cases, can even lead to the death of the patient. Therefore, it is compulsory to provide a model for finding a suitable treatment [2]. Machine learning is a powerful tool that uses previously recorded data as input to predict

future incidents [3]. Risk prediction plays a critical role in a clinical prescription for patients undergoing heart surgery. heart surgery departments of hospitals are producing remarkable recorded data during a day, which should be utilized by data scientists to quantify the patient's health and foresee future incidents [4]. The most important parameter after heart surgery is mortality because earlier diagnosis will raise the probability of patient's survival[5]. For this reason, different machine learning methods have been extended during past decades to predict mortality and, scholars have made myriad endeavors to enhance mortality prediction accuracy [6].

In 1999 J. Martinez-Alario et al. investigated and assessed the performance of general severity systems and compared them with the Parsonnet score to predict mortality after cardiac surgery[5]. In 2021 Dimitris Bertsimas et al. used a Decision tree for its interpretability and accuracy to predict mortality, postoperative mechanical ventilatory support time (MVST), and hospital length of stay (LOS) for patients who underwent Congenital Heart Surgery. They compared it to Logistic regression, Random Forest and Gradient Boosting. It turned out that Optimal classification trees outperform across all three models by reporting area under the curve[7]. There is a measure called the Euro score for predicting the mortality and probability of heart failure. Although most of the time it causes an overestimation, it continues to be used in the United Kingdom as regards a lack of alternative validated models. Machine learning models can be used to improve this overestimation. So, the Euro score is considered one of the essential features in the data set[4]. As it mentioned, Mortality prediction is a critical topic after surgery because with this information, many efforts can be made to save lives. Xia et al. [8] recommended an Artificial Neural Network (ANN) approach using recorded information in the first two days of an ICU to foresee the mortality risk state.

In 2000 Thomas G.Dietterich did an experimental comparison among three models, including bagging, boosting, and randomization, to improve the c4.5 algorithm performance. Moreover, his research's outcome indicates that in conditions with little or no classification noise, randomization can compete with bagging but is not as accurate as boosting. In

conditions with substantial classification noise, bagging has a better performance than boosting, and sometimes better than randomization. It motivated us to compare different ML models with our new model to enhance its' performance[9]. In addition to comparison, we wanted to know if machine learning can predict the mortality after heart surgery that Umberto Benedetto et al. have done it recently by the use of a neural network, random forest, naïve Bayes, and retrained LR based on features included in the EuroScore [4]. Also, some researches have been done by researchers about accurate diagnosis and effective treatment in heart disease with the application of ML. In 2017, Amin Poorieh et al. surveyed different ML methods and compared them, including Bayesian, KNN, SVM, Bagging, Boosting, and Stacking to predict heart disease[10].

Making better prediction accuracy of mortality can be crucial. Several researches have been done by scientists to prove that using ensemble methods can modify prediction outcomes. In 2017, Awad et al.[11] recommended an ensemble learning random forest(RF) and ended up that the proposed ensemble model has better prediction than other classification models. Ghose *et al.*[12] Moreover, Darabi *et al.* [13] reached similar outcomes by using the ensemble method to foresee the risk of mortality in ICU. These researches demonstrated that using a combination of classifiers can enhance prediction outcomes. Besides, according to Ghorbani and Ghosi's review paper[14] in the field of predictive models in medical diagnosis, it turned out that scholars had achieved better accuracy and prediction results while using ensemble approaches.

Other machine learning classifiers also are refined and compared with various models. One of the most efficient ensemble methods, according to research that has been done by Dehkordi and Sajedi in 2017, to predict depression in older adults, is the Stacking technique. This article's outcome, indicated that the proposed stacking model, which is a combination of K-Nearest-Neighbor, logistic regression, support vector machine, and decision tree, had higher accuracy in comparison to each of them[15]. Furthermore, this research aimed to compare different single and ensemble classifiers with the proposed new stacking model as regards Rosaida Rosly's research that shows a comparison of different models based on 10-fold cross-validation such as

three popular ensemble methods, which are boosting, bagging, and stacking for the combination and some single classifiers such as Naïve Bayes, Multi-Layer Perception and Decision Tree[16]. In 2019 Yoshihiko Raita et al. used some classification models, including Lasso regression, random forest, gradient boosted decision tree, and deep neural network to achieve clinical prediction outcomes, and then compared their performance[17].

Although the shortage of using ensemble models in predicting mortality after heart surgery should be considered, there are two other factors to improve prediction accuracy that should be managed, which are feature importance and solving unbalanced class distribution problems. As regards Muhammad Waqar et al.[18] used SMOTE to handle imbalanced data set problem for Heart Attack Prediction, which is a common problem especially in mortality prediction case, it is an efficient approach to handle this problem and is used in our research. Furthermore, our research shows that this problem can affect model performance[19]. Jale Bektas et al. worked on the classification of Real unbalanced Cardiovascular Data by Using Feature Selection and resampling Methods[20]. According to the importance of these factors, Roumani *et al.* [21] compared the performance of some prevalent machine learning approaches working on imbalanced data set problem. If the dataset has too many features, a high-dimensional problem may appear. In this situation, there are some features that are not essential and influential enough, so dimensionality reduction is required, as Anna Karen Garate et al.[22], had done in their research for heart disease prediction by PCA and random forest. Mohammad Al Khaldy and Chandrasekhar Kambhampati focused on different feature selection methods and resampling imbalanced class for heart failure data set[23].

There is an apparent lack of using ensemble methods and handling imbalanced data problems. A combination of some classifiers can lead to improving prediction results and accuracy compared to a simple classifier. This paper makes an effort to proposed a new ensemble model using the stacking approach to develop a robust early mortality forecasting model while solving highly unbalanced data problems by a combination of two resampling strategies.

There are some innovations and important processes in this paper as compared to similar researches:

- Predicting mortality after heart surgery by a new stacking model.
- Proposing a new ensemble classifier using the stacking approach.
- Solving the problem of highly imbalanced data set using a combination of SVM-SMOTE and Edited-Nearest-Neighbor approach.
- Applying both simple validation measures and 10-fold cross-validation methods to implement the validation test.
- Comparing the new stacking classifier with different single and ensemble machine learning classifiers.
- Measuring the performance of the proposed stacking model with different evaluation criteria such as accuracy, area under the ROC curve, Recall, Precision, and F1-Score.
- Using the Friedman test as a statistical test to analyze the differences among models and demonstrate the best one and validate the results.

The remnant of this article is as follows: section 2 explains all the data cleaning and preprocessing methods like handling imbalanced class problems and feature selection. Section 3 demonstrates the details about the new stacking ensemble model. In section 4, evaluation approaches are explained as a way to interpret the efficiency of the models. To show the acceptable performance of the recommended stacking model compared to other classifiers, Section 5 indicates the results and analysis. Eventually, section 6 shows the conclusions and proposes some directions for future researches.

2. MATERIAL & METHODS

Preprint does not need approval from a research ethics committee. A vital ingredient of the research, is problem perception. This research aims to develop an ensemble model using a stacking approach to predict mortality after heart surgery. It should be mentioned that all the coding process has been done in python, which is a constructive and beneficial language in the field of machine learning. In addition, all the practical experiments have been implemented

by a 2.40 GHz Intel Core i7 Lenovo with 16GB of RAM. The applied procedure to acquire the target of this research is portrayed in Figure 1.

This procedure includes, data understanding, data preprocessing, feature selection, stacking model usage, analyzing results.

Figure 1. Different steps of the applied procedure

2.1. DATA SET INFORMATION

In this paper, the chosen dataset is gathered and recorded manually from the hospitals related to Shahid Beheshti University of Medical Sciences and Health Services in Iran between 2015 and 2020. This data set is recorded during, and after a heart surgery(recovery) process that patient was hospitalized. Accordingly, the features are considered constant during the entire study. This data set encompasses 1632 records and 46 attributes after the data cleaning step. The number of 45 features are used in mortality prediction, and the other one is the response variable which is named mortality.

2.2. DATA PREPROCESSING

Data preprocessing is a necessary part of data mining and machine learning. The learning efficiency of a classifier depends markedly on the quality of the dataset. Consequently, preparing data is a critical step before entering it into a classifier[24]. It should be considered that the introduced data set had some missing data that it is handled after the data cleaning process. In this paper, Random Forest feature importance is used for feature selection. There are some more essential features among different features in a high dimensional data set that can be determined by the feature selection process in the data preprocessing step. This essential concept affects the machine learning model's performance[25]. By This technique, important and influential features can be ascertained to be used in prediction by machine learning models. Besides, feature selection approaches can lead to dimension reduction of a high dimensional data set by dropping the less effective or

noisy features so that the classifiers can be more efficient and precise. Feature selection techniques are divided into Wrapper and Filter techniques; which Wrapper method uses a classifier to select an optimal subset of features. In this paper, Random Forest feature importance is used according to its' popularity in medicine. It is a nonlinear technique that can be interpreted well. It provides feature importance measures by calculating the Gini importance[26]. In this paper, the selected data set had 45 features related to mortality prediction after heart surgery. After applying random forest feature importance, the importance of features is determined and listed in table 1. It should be mentioned that the first 18 features are given in the table because the other 27 features had less than 1 percent importance, and they had a slight effect in the performance of machine learning classifiers.

Recognizing the essential features can lead to dimension reduction, which can be helpful in mortality prediction. These 18 features predict Mortality status after heart surgery by different machine learning models and the new Stacking model.

2.2.1. IMBALANCED DATA PROBLEM

A balanced dataset is essential for creating an efficient training set. The unbalanced class distribution is a prevalent problem for real-world medical data and specifically mortality prediction. There is a minority class in predicting mortality because one of the two classes is significantly under-represented in the dataset [27]. Moreover, in this problem, the majority class highly dominates the minority class. Because of this reason, machine learning models tend to assign new observations to the majority class[28]. Consequently, this problem can cause poor performance of machine learning models in minority class prediction[29],[30].

It should be mentioned that the heart surgery dataset is markedly unbalanced. It comprises more samples from majority class that are survived patients (1583 cases of survival), while the minority class is too much smaller (only 49 cases of dead patients). According to this highly imbalanced dataset, it is imperative to handle this problem to help classifiers perform more efficiently and accurately. There are

Table 1. Main features of the studied heart surgery unit dataset with their importance using RF feature importance

Feature Name	Description	Type	Importance
Max-CR	Maximum amount of Cardiac rehabilitation	Numeric	12.9%
Euro score	A Criterion for the probability of heart attack	Numeric	10.5%
Day-Max-CR	Number of days spent on Max-CR	Numeric	10.1%
Weight	The weight of the patients in Kilograms	Numeric	7.8%
Day	The number of patient hospitalization days	Numeric	6.7%
Age	The age of the patients in years	Numeric	6.6%
Height	The height of the patients in cm	Numeric	6.5%
EF	The rate of heart's ejection fraction	Numeric	6.2%
D-Duration	Duration of surgery in hours	Numeric	3.9%
Diabetes	indicates the situation of patients whether if the patient has diabetes or not (0 for No and 1 for yes)	Nominal	2.1%
Injury	indicates the situation of patients whether if the patient has an injury or not (0 for No and 1 for yes)	Nominal	1.8%
HTN	indicates the situation of patients whether if the patient has hypertension or not (0 for No and 1 for yes)	Nominal	1.7%
Sex	The gender of patients (0 for men and 1 for women)	Nominal	1.6%
Renal failure	indicates the situation of patients whether if the patient has Renal failure or not (0 for No and 1 for yes)	Nominal	1.5%
AKI	indicates the situation of patients whether if the patient has Acute kidney injury or not (0 for No and 1 for yes)	Nominal	1.4%
Risk	indicates the situation of patients whether if the patient has surgery risk according to medic's opinion or not (0 for No and 1 for yes)	Nominal	1.3%
PH	indicates the situation of patients whether if the patient has Pulmonary hypertension according to medic's opinion or not (0 for No and 1 for yes)	Nominal	1.2%
Smoker	indicates the situation of patients whether if the patient is a smoker or not (0 for No and 1 for yes)	Nominal	1%

Source of data: heart surgery unit of hospitals related to Shahid Beheshti University of Medical Sciences and Health Services in Iran (2014-2020)

different sampling strategies which are built to handle the unbalanced data problem that can be implemented during the preprocessing step. Resampling is one of the most valuable methods to create a new dataset through increasing the minority class or decreasing the majority class[31]. This approach encompasses over-sampling and under-sampling strategies. Over-Sampling techniques increase the number of the minority class members in the training set, and the critical benefit of this technique is that no information from the original training set is lost because all members of both classes are

preserved[32],[33]. The synthetic minority Over-sampling technique (SMOTE) is one of the most prevalent and helpful over-sampling methods which can help to solve the unbalanced data problem[34]. The synthetic minority Over-sampling technique (SMOTE) is one of the most prevalent and helpful over-sampling strategies which can help to solve the unbalanced data problem[34]. This method produces synthetic data according to the feature space similarities among the minority class samples[35]. SMOTE generates artificial minority class examples to balance the dataset[36]. This paper tries to solve

unbalanced data problems using a combination of SVM-SMOTE over-sampling and Edited-Nearest-Neighbor under-sampling algorithms, which are extensively used in machine learning with imbalanced high-dimensional data that are increasingly used in medicine. SVM-SMOTE method has shown more acceptable performance than other resampling methods[37]. It helps predictive models to perform more accurately and reliably[38]. SMOTE resampling technique has different extended methods to handle imbalanced data problems. In 2020, Ghorbani and Ghousi compared them and the results of their research shows that the performance of SVM-SMOTE is better than others, and also, it can improve the performance of classifiers[39]. SVM-SMOTE concentrates on generating new minority class examples near borderlines with SVM to help establish boundaries among classes as regards data and density information which is crucial to synthesize minority classes. This paper combines the SVM-SMOTE with the Edited-Nearest-Neighbor method to prevail unbalanced data problems, and its' destructive influences on machine learning algorithms. It shows how a blend of resampling techniques, including over-sampling and under-sampling, can ameliorate the performance of a model[40].

K-Fold cross-validation technique is a helpful technique for data testing to evaluate the performance of classification models, and it determines how the statistical analysis outcomes are assigned to an independent dataset [41]. This paper uses random hold-out and shuffle 10-Fold cross-validation as two general forms of cross-validation. In the hold-out strategy, 75% of data is randomly assigned into the training set. The remaining 25% are assigned into the test set. Furthermore, in 10-Fold cross-validation, the original sample is randomly divided into ten equal size subsamples. Then a single subsample is kept as a test set, and the remaining subsamples are used as the training set. This method repeats this hold-out strategy ten times. In addition, the fact that all the resampling processes should be implemented on the training set is fundamental because otherwise, the synthetic observations can be seen by the classifier, and it can lead to overfitting and unreliable outcomes. It is genuinely compulsory that models only should be tested on unseen data. Therefore, the proposed combination of two resampling strategies including

SVM-SMOTE and Edited-Nearest-neighbor is only used on the training set while using hold-out and 10-fold cross-validation.

2.2.2. FEATURE SCALING

As most datasets in real-world contain markedly diverse features in sizes, units, and range, feature scaling comes up to standardize independent variables. This step is called normalization. This process should be done because the performance of many machine learning algorithms can be affected by feature scale diversity[42],[43]. Therefore, in this paper, features are rescaled; as a result, the entire used features have a standard normal distribution with $\mu=0$ and $\sigma=1$ where μ is the average, and σ is the standard deviation. The formula of this type of rescaling is given below:

$$Z = \frac{X - \mu}{\sigma} \quad (1)$$

3. MACHINE LEARNING ALGORITHMS

Many machine learning algorithms have been developed to solve various classification problems in different fields of science[44], such as Logistic Regression [31], Naive Bayes [45], Decision Tree[46], [47], Random Forest[48], K-Nearest-Neighbor[49], Gradient Boosting(XG-boost)[50], [51], Artificial Neural Network[52] and etc. These algorithms are applied after preprocessing process when the dataset is ready to be learned by the computer. In this paper, some of them are used and compared with the Stacking model. All of these machine learning models are given in Table 2. with their efficient parameters.

3.1. NEW STACKING MODEL

This research proposed a new combination of models based on the stacking algorithm, which uses a meta-model to combine predictions from contributing members. This new model is a combination of five single and ensemble classifiers, developed to improve the predicting performance of mortality after hurt surgery. Stacking is an efficient ensemble technique because it can balance bias and variance to reduce the total error and achieve better prediction. Ensemble machine learning classifiers aim is to reduce bias and

Table 2. Machine learning models with their specific parameters' settings.

methods	parameters
Logistic Regression	C=13, penalty=l2
Decision Tree	Criterion=entropy, Splitter=best
Random Forest	N_Estimator =100, Min_Samples_Leaf=1
XG-Boost	N_Estimator=75, Learning_rate=1, Loss=deviance
Bagging	-
Naïve Bayes	-
K-Nearest-Neighbor	N-Neighbors=8

variance of classifiers to build a robust classifier that can achieve better performance. In the stacking method, the output of the base classifiers (Level 0) will be used as training data for another classifier named Meta-classifier (Level1) to estimate the same target function. The target of stacking is combining various strong sets of models. In this paper, the base classifiers of the proposed stacking model are Logistic Regression, Bagging, Decision Tree, Random Forest, and XG-Boost. The Meta-classifier is Logistic Regression which is selected according to their better performance rather than other classifiers. Although the original classifiers predict the minority class weakly, they work efficiently for some parts of the data. So, each model works as a booster to sharpen the efficiency of the ensemble model[53]. In the stacking ensemble approach, Meta-learner tries to discover how best to combine the output of the base learners. Figure 2 demonstrates the levels and outputs of the proposed stacking model.

Figure 2. proposed Stacking ensemble model.

As it is seen, this paper links two simple classifiers, including Logistic Regression and Decision Tree, as the best single classifiers with Random Forest, XG-

Boost, and Bagging as the three best ensemble classifiers to build a powerful ensemble machine learning model.

4. EVALUATION METHODS

The evaluating step of machine learning algorithms is an inseparable ingredient of applying classification models. Various measures can be used to assess the performance and validation of machine learning models. In this paper, Accuracy, Area under the ROC curve (AUC), Recall, Precision, and F1-Score are used as metric measurement systems. Besides, a statistical significance test is implemented to survey the differences among different models.

4.1. STATISTICAL EVALUATION

Evaluation methods are the tools for comparing different classification methods. In this research, according to the Shapiro normality test, which is used to survey the normality of data, it turns out that the data does not have a normal distribution. Because the p-value of this test is less than α ($\alpha=0.05$), and the null hypothesis (data is normal) will be rejected. Therefore, in this case, the Friedman test, which is a nonparametric equivalent of the repeated-measures ANOVA, can be used to survey the differences among machine learning models[54]. The null hypothesis of this test is that the performance of all models is similar and rejection of it demonstrates that one or more of the paired Classifiers perform differently. This paper uses the accuracy obtained by 10-fold cross-validation. The Friedman test ranks the data of each fold together; then there are some values of ranks for each model[55]. As a result, this test provides a sum of ranks for each model that can lead to detecting the most effective classifiers among others.

5. RESULTS & DISCUSSION

5.1. MACHINE LEARNING MODEL RESULTS

There are various classification models. In this paper, the most common and useful models are compared with the proposed stacking ensemble model. Most of the assessment criteria indicate that the Stacking model had better performance and improved the ability of mortality prediction after heart surgery. Model validation procedures used in this research are based on the random hold-out and shuffle 10-fold cross-validation techniques. Table 3 demonstrates the performance of the different models and the introduced stacking model using the hold-out strategy. Besides, for the better interpretation of the

Table 3. Performance of the models based on the 75/25 random hold-out strategy

Model	Test set accuracy	AUC	Recall	Precision	F1-Score Macro-average
Proposed Stacking model	97.30%	93.10%	71.00%	86.50%	77.00%
Random Forest	97.05%	91.80%	69.50%	82.50%	69.00%
Logistic Regression	96.81%	91.60%	74.00%	71.00%	72.00%
Bagging classifier	96.07%	80.70%	63.50%	68.50%	65.50%
XG-Boost	94.85%	76.50%	59.00%	60.00%	59.50%
Decision Tree	94.61%	62.80%	63.00%	61.00%	62.00%
K-Nearest-Neighbor	93.62%	62.40%	55.00%	54.50%	55.00%
Naïve bayes	92.65%	76.50%	68.50%	59.50%	62.50%

Figure 3. The test accuracy results for models based on the 75/25 random hold-out strategy.

differences among the model performances, Figure 3 is prepared to compare the test accuracy results.

5.1.1. HOLD-OUT METHOD RESULTS

In this paper, the hold-out method assigns 75% of the dataset to the training set. The remnant 25% of the data is assigned to the test set. In addition, as mentioned, a combination of SVM-SMOTE as an over-sample strategy and Edited-Nearest-Neighbor as an under-sample strategy is introduced to solve the highly unbalanced data problem in this case. Accuracy is the most prevalent evaluation criterion to show every classifier's performance. Although it is easy to understand, it is not enough to correctly judge the model because many critical factors should be

noticed in evaluating the performance of a model. According to Table 4 results, it turns out that the proposed stacking model performs well based on test set accuracy and has improved the accuracy compared to other models. Furthermore, the introduced combination of resampling methods solved the unbalanced data problem efficiently, so the accuracy is more reliable. Because it works with ones and zeros, so there is uncertainty about quantifying a probability value. AUC is a crucial evaluation metric that indicates how much a model is efficient in distinguishing between the two classes. Therefore, the higher AUC reveals the better capability of this issue. Stacking model with 93.10% of the AUC metric percentage is the best classifier among the other models. This result means that the introduced stacking model has the ability to distinguish between survival and dead patients with a 93.10% of chance which is better than other classifiers, even Random Forest and Logistic Regression. Figure 4 shows a multiple ROC curve which indicates the mentioned outperforming of the stacking model compared to base and meta classifiers.

The other proper measures to evaluate the models' efficiency are Precision and Recall. Precision indicates what proportion of positive identifications was correct, while Recall shows what proportion of actual positive was appropriately identified. Although, Precision and Recall often have a trade-off, both of them should be considered in model evaluation. Most of the time, improving precision reduces Recall[56]. The Stacking model has improved Recall and Precision together. It achieved 71% with the Recall on average, which is 3% percent less than the best Recall (Logistic Regression), and 86.5% with the Precision on average, which is the best number among all of the models. The Stacking model identifies 71% of survival and dead patients on

average correctly, and when it ascertains a patient as survival or dead, it is correct 86.5% of the time on average. It should be mentioned that the macro average is used to report the models' performance results because when the dataset was imbalanced, models did not predict well the minority class. Therefore, macro average is used, which results in a bigger penalization and can reveal more fair judgement. The F1-score metric is a measure that can make a more fairly Analysis and comparison among the models by the Recall and the Precision since it is the harmonic average of both metrics. It ascertains how accurate and authoritative is the prediction of the model. In this paper, the macro average of the introduced stacking model achieved the most F1-score among the used models. These results indicate the desirable efficiency of the stacking model. Using Random Forest feature importance shows the effect and importance of each feature in predicting mortality and survival of patients.

Figure 4. The multiple ROC curve diagram.

The results demonstrate that it is not necessary to record 27 other medical features, and time can be optimized. The maximum amount of Cardiac rehabilitation (Max-CR), Euro score, and Day-Max-CR are the most essential features. So, it turns out the Euro score as a criterion of heart attack probability and cardiac rehabilitation should be more noticed. As was expected, Age is vital to determine the chance of survival or death after heart surgery.

The information about these features is provided in Table1.

5.1.2. K-FOLD CROSS VALIDATION METHOD RESULTS

k-fold cross-validation is another method of assessing validation[57]. Shuffle 10-fold cross-validation is used in this research, which splits the dataset into ten subsets. The results of applied 10-fold cross-validation results using machine learning models are shown in Table 4, which indicates the obtained accuracy of each fold by different models. According to results shown in Table 4, the introduced stacking model achieved the highest average of 10 folds accuracy with low variance, 96.97%, and it had a slight difference with the hold-out strategy, which means its' accuracy is reliable, and the model performed acceptably. The accuracy resulted from 1st fold, 96.88% was the lowest. The highest accuracy is related to the third fold, 97.03%. Random Forest model achieved the second-highest accuracy average, 96.87, and according to the whole assessments and validations results, it is the latter most well-performed model compared to others. It is better to be mentioned that the proposed stacking model was the best in almost all validation and assessing metrics. It achieved the best rank in all of them except in Recall that achieved the second stage after the LR model. In addition, Figure 5 shows the comparison among machine learning classifiers while using 10-fold cross-validation and hold-out strategies. The difference between Naïve Bayes results by using the two strategies indicates this model's hold-out accuracy is not reliable, and it may not perform as well as it seemed. Other models' 10-fold cv accuracy emphasize their hold-out accuracy, especially stacking, Random Forest, and Logistic Regression models. Cross-validation is often the preferred method because it gives a better demonstration of how well the model will perform on unseen data.

5.1.3 STATISTICAL TEST RESULTS

Statistical significance tests can be used as one of the valuable ways to select the best model. The first assumption of the ANOVA test is that the 10-fold cv samples are drawn from a normal distribution. This paper's assumption of the ANOVA test is violated because the Shapiro statistical normality test reached a p-value=0, which is less than $\alpha=0.05$. Therefore, the null hypothesis which is the normality of data is rejected, and the ANOVA test can't be used. Table 5 represents the result of the Shapiro test. The Friedman test is suitable to compare machine learning classifiers in this case. Table 6 indicates the results of this test. The null hypothesis of this test is rejected because the p-value=0 is less than the significance level of 0.05. It means that at least one

Table 4. Performance of the models based on the shuffle 10-fold cross-validation

Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10	Average
Stacking	96.88%	96.96%	97.03%	96.96%	96.97%	97.00%	96.95%	96.97%	97.02%	97.01%	96.97%
RF	97.31%	96.94%	96.94%	96.90%	96.88%	96.77%	96.81%	96.77%	96.81%	96.76%	96.87%
LR	96.50%	96.64%	96.20%	96.58%	96.49%	96.94%	96.84%	96.82%	96.74%	96.86%	96.66%
Bagging	96.33%	96.45%	96.24%	96.14%	96.18%	96.32%	96.35%	96.38%	96.49%	96.30%	96.31%
NB	88.87%	88.87%	88.85%	88.74%	88.69%	88.69%	88.69%	88.65%	88.70%	88.71%	88.74%
DT	95.82%	95.20%	94.59%	94.89%	94.63%	94.82%	94.13%	94.52%	94.37%	94.85%	94.78%
XG-boost	96.33%	96.20%	95.76%	95.84%	95.98%	96.03%	96.04%	96.07%	95.91%	95.95%	95.81%
KNN	96.55%	96.55%	96.55%	96.55%	96.55%	96.56%	96.56%	96.56%	96.56%	96.56%	96.56%

FIGURE 5. Comparing shuffle 10-fold cross-validation and hold-out accuracy results.

of the classifiers performs differently. Table 7 demonstrates the median and sum of ranks achieved by the Friedman test. The median indicates the midpoint value, which is the point where half the data points are above it and the other half is the other way round. In addition, the median of all data points which is shown as the overall median. Additionally, the median response for the stacking

TABLE 5. Shapiro normality test result

	Mean	Standard Deviation	statistics	p-value
Tested data	50.3%	4.43%	50.8%	0.00

TABLE 6. The Friedman test results.

Friedman test	Degrees of freedom	chi-square	p-value
	7	135	0.00

TABLE 7. Additional information from Friedman test results.

Rank	Classifiers	Median	Sum of Ranks
1	Logistic Regression	97.05	257
2	Stacking model	97	221.5
3	Random Forest	97	210
4	XG-Boost	96.99	173.5
5	K-Nearest-Neighbor	96.93	159
6	Bagging classifier	96.53	96
7	Decision Tree	94.82	60
8	Naïve Bayes	89.89	30
	Overall	96.99	

model is higher than the overall median. The result of the sum of Ranks shows that the stacking model is better than all classifiers except Logistic Regression.

6. CONCLUSION

Nowadays, a vast amount of data is collected and produced by the development of health care systems and biomedical equipment. Data processing and releasing valuable patterns and information is a way to save many lives. Predicting mortality after heart surgery is the essential concepts in medical data mining. An accurate prediction of the mortality status of patients waiting for heart surgery can give helpful information to save lives and lessen costs and time; so, it is compulsory to utilize it in patients as soon as possible. This paper tries to propose a stacking ensemble model, which uses features selected by Random Forest feature importance technique to provide an early mortality prediction while solving the unbalanced data problem by using a combination of SVM-SMOTE over-sampling strategy and Edited-Nearest-Neighbor under-sampling technique. The random hold-out and shuffle 10-fold cross-validation are used as two validation procedures to assess the machine learning models' stability and performance. Moreover, a statistical test called the Friedman test is used as another measure of performance. After solving the highly unbalanced data problem, well-known models and the introduced stacking ensemble model are applied using the random hold-out method on balanced data. According to the assessment outcomes, the introduced stacking model has acceptable performance and outperforms all other models using various assessment metrics. It should be mentioned that Random Forest and Logistic Regression have an acceptable performance as well. The evaluation results of machine learning models obtained using the shuffle 10-fold cross-validation method indicate similar results to choose the best model. The introduced stacking ensemble model achieved higher accuracy than other models with a low figure of acceptable variance. Using the stacking model, which is a combination of some single and ensemble classifiers, improved the mortality prediction ability after heart surgery according to different validation and assessment techniques. The stacking model achieved the best performance in all the measures except Recall. In Recall, it achieved the

second stage. Therefore, it seems that the new model had the best performance. It should be considered that RF and LR were the two other efficient models after the stacking model. Many ways are existed to refine and sharpen this research. Future researches can be done to make more powerful models. Also, other combinations of resampling techniques can be implemented to handle imbalanced data problems.

REFERENCES

- [1] I. R. Bardhan and M. F. Thouin, "Health information technology and its impact on the quality and cost of healthcare delivery," *Decis. Support Syst.*, vol. 55, no. 2, pp. 438–449, 2013, doi: 10.1016/j.dss.2012.10.003.
- [2] S. Ratnakar, K. Rajeswari, and R. Jacob, "Prediction of Heart Disease Using Genetic Algorithm," *Int. J. Adv. Comput. Eng. Networking*, vol. 1, no. 2, pp. 51–55, 2013.
- [3] A. Banerjee *et al.*, "Machine learning for subtype definition and risk prediction in heart failure, acute coronary syndromes and atrial fibrillation: systematic review of validity and clinical utility," *BMC Med.*, vol. 19, no. 1, pp. 1–14, 2021, doi: 10.1186/s12916-021-01940-7.
- [4] L. Myslivecek and U. . Benedetto, "Machine learning for risk prediction of in-hospital mortality in adult cardiac surgery," *Eur. Heart J.*, vol. 41, no. Supplement_2, p. 2668, 2020, doi: 10.1093/ehjci/ehaa946.2668.
- [5] I. D. Tuesta, E. Plasencia, M. Santana, and M. L. Mora, "Comparative Performance of Parsonnet and General Severity Systems," pp. 2378–2383, 1999.
- [6] A. Awad, "A Data Mining Approach for Early Mortality Prediction of Patients in Intensive Care Units Aya Awad," *Researchportal.Port.Ac.Uk*, 2016, [Online]. Available: https://researchportal.port.ac.uk/portal/files/12937925/P_RISMA_A_DM_Approach_for_EMP_of_Patient_in_ICUs.pdf.
- [7] D. Bertsimas *et al.*, "Adverse Outcomes Prediction for Congenital Heart Surgery : A Machine Learning Approach," 2021, doi: 10.1177/21501351211007106.
- [8] H. Xia, B. J. Daley, A. Petrie, and X. Zhao, "A neural network model for mortality prediction in ICU," *Comput. Cardiol. (2010).*, vol. 39, pp. 261–264, 2012.
- [9] T. G. Dietterich, "Experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization," *Mach. Learn.*, vol. 40, no. 2, pp. 139–157, 2000, doi: 10.1023/A:1007607513941.
- [10] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia, and J. Gutierrez, "A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease," *Proc. - IEEE Symp. Comput. Commun.*, pp. 204–207, 2017, doi: 10.1109/ISCC.2017.8024530.
- [11] A. Awad, M. Bader-El-Den, J. McNicholas, and J. Briggs, "Early hospital mortality prediction of intensive care unit patients using an ensemble learning approach,"

- Int. J. Med. Inform.*, vol. 108, pp. 185–195, 2017, doi: 10.1016/j.ijmedinf.2017.10.002.
- [12] S. Ghose, J. Mitra, S. Khanna, and J. Dowling, “An Improved Patient-Specific Mortality Risk Prediction in ICU in a Random Forest Classification Framework,” *Stud. Health Technol. Inform.*, vol. 214, pp. 56–61, 2015, doi: 10.3233/978-1-61499-558-6-56.
- [13] H. R. Darabi, D. Tsinis, K. Zecchini, W. F. Whitcomb, and A. Liss, “Forecasting mortality risk for patients admitted to intensive care units using machine learning,” *Procedia Comput. Sci.*, vol. 140, pp. 306–313, 2018, doi: 10.1016/j.procs.2018.10.313.
- [14] R. Ghorbani and R. Ghousi, “Predictive data mining approaches in medical diagnosis: A review of some diseases prediction,” *Int. J. Data Netw. Sci.*, vol. 3, no. 2, pp. 47–70, 2019, doi: 10.5267/j.ijdns.2019.1.003.
- [15] S. K. Dehkordi and H. Sajedi, “A prescription-based automatic medical diagnosis system using a stacking method,” *SISY 2017 - IEEE 15th Int. Symp. Intell. Syst. Informatics, Proc.*, pp. 187–191, 2017, doi: 10.1109/SISY.2017.8080550.
- [16] R. Rosly, M. Makhtar, M. K. Awang, M. I. Awang, and M. N. A. Rahman, “Analyzing performance of classifiers for medical datasets,” *Int. J. Eng. Technol.*, vol. 7, no. 2, pp. 136–138, 2018, doi: 10.14419/ijet.v7i2.15.11370.
- [17] Y. Raita, T. Goto, M. K. Faridi, D. F. M. Brown, C. A. Camargo, and K. Hasegawa, “Emergency department triage prediction of clinical outcomes using machine learning models,” *Crit. Care*, vol. 23, no. 1, pp. 1–13, 2019, doi: 10.1186/s13054-019-2351-7.
- [18] M. Waqar, H. Dawood, H. Dawood, N. Majeed, A. Banjar, and R. Alharbey, “An Efficient SMOTE-Based Deep Learning Model for Heart Attack Prediction,” *Sci. Program.*, vol. 2021, 2021, doi: 10.1155/2021/6621622.
- [19] S. Fotouhi, S. Asadi, and M. W. Kattan, “A comprehensive data level analysis for cancer diagnosis on imbalanced data,” *J. Biomed. Inform.*, vol. 90, no. January, p. 103089, 2019, doi: 10.1016/j.jbi.2018.12.003.
- [20] J. Bektas, T. Ibricki, and I. T. Ozcan, “Classification of Real Imbalanced Cardiovascular Data Using Feature Selection and Sampling Methods: A Case Study with Neural Networks and Logistic Regression,” *Int. J. Artif. Intell. Tools*, vol. 26, no. 6, pp. 1–12, 2017, doi: 10.1142/S0218213017500191.
- [21] Y. F. Roumani, J. H. May, D. P. Strum, and L. G. Vargas, “Classifying highly imbalanced ICU data,” *Health Care Manag. Sci.*, vol. 16, no. 2, pp. 119–128, 2013, doi: 10.1007/s10729-012-9216-9.
- [22] A. K. Gárate-Escamila, A. Hajjam El Hassani, and E. Andrés, “Classification models for heart disease prediction using feature selection and PCA,” *Informatics Med. Unlocked*, vol. 19, 2020, doi: 10.1016/j.imu.2020.100330.
- [23] M. Al Khaldy, “Resampling Imbalanced Class and the Effectiveness of Feature Selection Methods for Heart Failure Dataset,” *Int. Robot. Autom. J.*, vol. 4, no. 1, pp. 37–45, 2018, doi: 10.15406/iratj.2018.04.00090.
- [24] S. Zhang, C. Zhang, and Q. Yang, “Data preparation for data mining,” *Appl. Artif. Intell.*, vol. 17, no. 5–6, pp. 375–381, 2003, doi: 10.1080/713827180.
- [25] R. C. Chen, C. Dewi, S. W. Huang, and R. E. Caraka, “Selecting critical features for data classification based on machine learning methods,” *J. Big Data*, vol. 7, no. 1, 2020, doi: 10.1186/s40537-020-00327-4.
- [26] M. Saarela and S. Jauhiainen, “Comparison of feature importance measures as explanations for classification models,” *SN Appl. Sci.*, vol. 3, no. 2, pp. 1–12, 2021, doi: 10.1007/s42452-021-04148-9.
- [27] S. Tripathi, S. Batra, and S. Pandey, “Unbiased Mortality Prediction for Unbalanced Data Using Machine Learning,” *Proc. - 2019 Int. Conf. Electr. Electron. Comput. Eng. UPCON 2019*, pp. 3–7, 2019, doi: 10.1109/UPCON47278.2019.8980003.
- [28] T. Liu, W. Fan, and C. Wu, “A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset,” *Artif. Intell. Med.*, vol. 101, p. 101723, 2019, doi: 10.1016/j.artmed.2019.101723.
- [29] J. Liu *et al.*, “Mortality prediction based on imbalanced high-dimensional ICU big data,” *Comput. Ind.*, vol. 98, pp. 218–225, 2018, doi: 10.1016/j.compind.2018.01.017.
- [30] D. Scrutinio *et al.*, “Machine learning to predict mortality after rehabilitation among patients with severe stroke,” *Sci. Rep.*, vol. 10, no. 1, pp. 1–10, 2020, doi: 10.1038/s41598-020-77243-3.
- [31] D. R. Cox, “The Regression Analysis of Binary Sequences Published by : Wiley for the Royal Statistical Society Stable URL : <http://www.jstor.org/stable/2983890>,” *J. R. Stat. Soc.*, vol. 20, no. 2, pp. 215–242, 1958.
- [32] M. M. Rahman and D. N. Davis, “Addressing the Class Imbalance Problem in Medical Datasets,” *Int. J. Mach. Learn. Comput.*, no. April 2013, pp. 224–228, 2013, doi: 10.7763/ijmlc.2013.v3.307.
- [33] J. A. Sáez, J. Luengo, J. Stefanowski, and F. Herrera, “SMOTE-IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering,” *Inf. Sci. (Nij)*, vol. 291, no. C, pp. 184–203, 2015, doi: 10.1016/j.ins.2014.08.051.
- [34] M. Cengiz Colak, E. Karaaslan, C. Colak, A. K. Arslan, and N. Erdil, “Handling imbalanced class problem for the prediction of atrial fibrillation in obese patient,” *Biomed. Res.*, vol. 28, no. 7, pp. 3293–3299, 2017.
- [35] D. Devi, S. kr Biswas, and B. Purkayastha, “Redundancy-driven modified Tomek-link based undersampling: A solution to class imbalance,” *Pattern Recognit. Lett.*, vol. 93, pp. 1339–1351, 2017, doi: 10.1016/j.patrec.2016.10.006.
- [36] H. Ishwaran and R. O’Brien, “Commentary: The problem of class imbalance in biomedical data,” *J. Thorac. Cardiovasc. Surg.*, vol. 161, no. 6, pp. 1940–1941, 2021, doi: 10.1016/j.jtcvs.2020.06.052.
- [37] M. Alghamdi, M. Al-Mallah, S. Keteyian, C. Brawner, J. Ehrman, and S. Sakr, “Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford Exercise Testing (FIT) project,” *PLoS One*, vol. 12, no. 7, pp. 1–15, 2017, doi:

- 10.1371/journal.pone.0179805.
- [38] Y. Liu, A. An, and X. Huang, "Boosting prediction accuracy on imbalanced datasets with SVM ensembles," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 3918 LNAI, pp. 107–118, 2006, doi: 10.1007/11731139_15.
- [39] R. Ghorbani and R. Ghousi, "Comparing Different Resampling Methods in Predicting Students' Performance Using Machine Learning Techniques," *IEEE Access*, vol. 8, pp. 67899–67911, 2020, doi: 10.1109/ACCESS.2020.2986809.
- [40] J. V. Hansen, "Combining predictors: comparison of five meta machine learning methods," *Inf. Sci. (Nijl.)*, vol. 119, no. 1–2, pp. 91–105, 1999, doi: 10.1016/S0020-0255(99)00052-3.
- [41] D. Normawati and D. P. Ismi, "K-Fold Cross Validation for Selection of Cardiovascular Disease Diagnosis Features by Applying Rule-Based Datamining," *Signal Image Process. Lett.*, vol. 1, no. 2, pp. 23–35, 2019, doi: 10.31763/simple.v1i2.3.
- [42] S. Aksoy and R. M. Haralick, "Feature normalization and likelihood-based similarity measures for image retrieval," *Pattern Recognit. Lett.*, vol. 22, no. 5, pp. 563–582, 2001, doi: 10.1016/S0167-8655(00)00112-4.
- [43] M. Ebrahimi, M. Mohammadi-Dehcheshmeh, E. Ebrahimi, and K. R. Petrovski, "Comprehensive analysis of machine learning models for prediction of sub-clinical mastitis: Deep Learning and Gradient-Boosted Trees outperform other models," *Comput. Biol. Med.*, vol. 114, no. September, p. 103456, 2019, doi: 10.1016/j.combiomed.2019.103456.
- [44] K. Y. Ngiam and I. W. Khor, "Big data and machine learning algorithms for health-care delivery," *Lancet Oncol.*, vol. 20, no. 5, pp. e262–e273, 2019, doi: 10.1016/S1470-2045(19)30149-4.
- [45] S. Kumar, A. Jain, and P. Mahalakshmi, "Enhancement of Healthcare Using Naïve Bayes Algorithm and Intelligent Datamining of Social Media," *Int. J. Appl. Eng. Res.*, vol. 13, no. 6, pp. 4109–4112, 2018, [Online]. Available: <http://www.ripublication.com>.
- [46] S. R. Safavian and D. Landgrebe, "A Survey of Decision Tree Classifier Methodology," *IEEE Trans. Syst. Man Cybern.*, vol. 21, no. 3, pp. 660–674, 1991, doi: 10.1109/21.97458.
- [47] W. Du, W. Du, Z. Zhan, and Z. Zhan, "Building decision tree classifier on private data," *Proc. IEEE Int. Conf. Privacy, Secur. data mining-Volume 14*, pp. 1–8, 2002, [Online]. Available: <http://portal.acm.org/citation.cfm?id=850784>.
- [48] A. M. Prasad, L. R. Iverson, and A. Liaw, "Newer classification and regression tree techniques: Bagging and random forests for ecological prediction," *Ecosystems*, vol. 9, no. 2, pp. 181–199, 2006, doi: 10.1007/s10021-005-0054-1.
- [49] R. K. Nowicki, "Rough Nearest Neighbour Classifier," *Stud. Comput. Intell.*, vol. 802, pp. 133–159, 2019, doi: 10.1007/978-3-030-03895-3_6.
- [50] A. Kilic *et al.*, "Predictive Utility of a Machine Learning Algorithm in Estimating Mortality Risk in Cardiac Surgery," *Ann. Thorac. Surg.*, vol. 109, no. 6, pp. 1811–1819, 2020, doi: 10.1016/j.athoracsur.2019.09.049.
- [51] E. Bauer and R. Kohavi, "Empirical comparison of voting classification algorithms: bagging, boosting, and variants," *Mach. Learn.*, vol. 36, no. 1, pp. 105–139, 1999, doi: 10.1023/a:1007515423169.
- [52] A. K. Jain, J. Mao, and K. M. Mohiuddin, "Artificial neural networks: A tutorial," *Computer (Long. Beach. Calif.)*, vol. 29, no. 3, pp. 31–44, 1996, doi: 10.1109/2.485891.
- [53] I. D. Mienye, Y. Sun, and Z. Wang, "An improved ensemble learning approach for the prediction of heart disease risk," *Informatics Med. Unlocked*, vol. 20, p. 100402, 2020, doi: 10.1016/j.imu.2020.100402.
- [54] M. Friedman, "The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance," *J. Am. Stat. Assoc.*, vol. 32, no. 200, pp. 675–701, 1937, doi: 10.1080/01621459.1937.10503522.
- [55] M. Friedman, "A Comparison of Alternative Tests of Significance for the Problem of m Rankings," *Ann. Math. Stat.*, vol. 11, no. 1, pp. 86–92, 1940, doi: 10.1214/aoms/1177731944.
- [56] C. W. Cleverdon, "Journal of documentation," *J. Doc.*, vol. 22, no. 3, pp. 266–268, 1966, doi: 10.1108/eb026404.
- [57] D. Gu, K. Su, and H. Zhao, "A case-based ensemble learning system for explainable breast cancer recurrence prediction," *Artif. Intell. Med.*, vol. 107, no. August 2019, p. 101858, 2020, doi: 10.1016/j.artmed.2020.101858.