

Received July 14, 2020, accepted July 29, 2020, date of publication July 31, 2020, date of current version August 13, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3013320

A New Hybrid Predictive Model to Predict the Early Mortality Risk in Intensive Care Units on a Highly Imbalanced Dataset

RAMIN GHORBANI¹, ROUZBEH GHOUSI², AHMAD MAKUI², AND ALIREZA ATASHI^{3,4}

¹Delft University of Technology, 2600 GA Delft, The Netherlands

²School of Industrial Engineering, Iran University of Science and Technology, Tehran 16846-13114, Iran

³e-Health Department, Virtual School, Tehran University of Medical Sciences, Tehran 19617-33114, Iran

⁴Cancer Informatics Research Group, Clinical Research Department, Breast Cancer Research Center, Motamed Cancer Institute, ACECR, Tehran 14155-4364, Iran

Corresponding author: Rouzbeh Ghousi (ghousi@iust.ac.ir)

ABSTRACT Due to the development of biomedical equipment and healthcare level, especially in the Intensive Care Unit (ICU), a considerable amount of data has been collected for analysis. Mortality prediction in the ICUs is considered as one of the most important topics in the healthcare data analysis section. A precise prediction of the mortality risk for patients in ICU could provide us with valuable information about patients' lives and reduce costs at the earliest possible stage. This paper aims to introduce a new hybrid predictive model using the Genetic Algorithm as a feature selection method and a new ensemble classifier based on the combination of Stacking and Boosting ensemble methods to create an early mortality prediction model on a highly imbalanced dataset. The SVM-SMOTE method is used to solve the imbalanced data problem. This paper compares the new model with various machine learning models to validate the efficiency of the introduced model. The achieved results using the shuffle 5-fold cross-validation and random hold-out methods indicate that the new hybrid model has the best performance among other classifiers. Additionally, the Friedman test is applied as a statistical significance test to examine the differences between classifiers. The results of the statistical analysis prove that the proposed model is more effective than other classifiers. Furthermore, the proposed model is compared to APACHE and SAPS scoring systems and is benchmarked against state-of-the-art predictive models applied to the MIMIC dataset for experimental validation and achieved promising results as it outperformed the state-of-the-art models.

INDEX TERMS Classification, hybrid predictive model, stacking ensemble method, boosting ensemble method, intensive care unit (ICU), imbalanced data problem, machine learning in healthcare, SVM-SMOTE method, Friedman test.


I. INTRODUCTION

Technological advancements in healthcare have saved countless lives and improved the quality of living. Not only has technology changed patients' experiences from a cure, but it has also had a significant influence on medical diagnosis [1]. The healthcare industry understands the potential of using machine learning in healthcare to dramatically develop administrative functions, clinical decision making, disease diagnosis, and patient monitoring. Determining the wrong treatment for patients not only waste time and money but also can cause unfavorable consequences such as a patient's death. Accordingly, it is necessary to have a system for diagnosing

and choosing the proper treatment [2]. Machine learning is the experimental study of analytical models and algorithms that builds a mathematical model of sample data to present predictions or decisions [3], [4].

Intensive Care Unit (ICU) is a special section of a hospital or healthcare department that provides intensive treatment medicine. ICU is one of the most critical operating environments in a hospital. Patients will be transferred to an ICU when it is clear that their conditions require constant and comprehensive monitoring and adjustment [5]. ICUs generate a considerable amount of data every day, which will be used to quantify the patient's health and predict future outcomes [6].

One of the essential issues in the ICU is patient mortality. It is well known that the mortality prediction is so important, and earlier detection and diagnosis will increase the

The associate editor coordinating the review of this manuscript and approving it for publication was Hong-Mei Zhang .

possibility of saving patients' life [7], [8]. That is why over the past few decades, different machine learning models and numerous scoring systems such as APACHE [9], SAPS [10], MPM [11], and SOFA [12] have been extended to predict mortality, and researchers have been dedicated themselves to the study of enhancing the mortality prediction accuracy and assessing mortality risk in ICU patients [13].

In 2012, the mortality prediction of ICU patients became the topic of the Physionet challenge to increase the development of new machine learning algorithms [14]. In response to the Physionet challenge, Xia *et al.* [15] introduced an Artificial Neural Network (ANN) model using collected data during the first two days of an ICU to predict the risk of mortality. Also, Dybowski *et al.* [16] developed an ANN algorithm optimized by a Genetic Algorithm (GA) that could be implemented in intensive care units. This research compared the ANN algorithm with Logistic Regression (LR) and reported the better performance of the ANN algorithm over the LR while achieving 86% with AUC. However, some research such works as Doig *et al.* [17], Clermont *et al.* [18], and Silva *et al.* [19] compared LR and ANN models and concluded that these two algorithms perform almost similarly in the prediction of mortality.

Other machine learning models have also been developed and compared with different models. Moridani *et al.* [20] developed a Support Vector Machine (SVM) algorithm to predict mortality risk of the cardiovascular patients in ICU and concluded that SVM achieves better results than the ANNs algorithm. Luaces *et al.* [21] presented a machine learning method based on the SVM and reported a comparison between this method and LR. Furthermore, Houthoof *et al.* [22] applied different machine learning models to predict patient mortality and length of stay (LOS) in ICU. The results informed that SVM achieves the best outcomes in terms of patient mortality prediction. Nevertheless, Kim *et al.* [23] assessed the performance of different data mining techniques, and the results revealed that the Decision Tree (DT) algorithm slightly outperforms the other data mining techniques.

It is a well-established fact that it can be challenging to improve the prediction accuracy of a model. Ensemble modeling is one of the most significant ways to improve the performance of a model. In 2017, Awad *et al.* [13] proposed an ensemble learning Random Forest (RF) and concluded that the introduced ensemble model outperforms other prediction models. Additionally, Ghose *et al.* [24] and Darabi *et al.* [25] achieved the same results by using ensemble models to predict mortality risk in ICU. These studies proved that using ensemble models can improve prediction results. Moreover, Ghorbani and Ghousi [2] reviewed the predictive data mining approaches in medical diagnosis, and the results declared that researchers had obtained better prediction accuracy while using ensemble models.

The lack of using hybrid and ensemble models in predicting mortality risk within the ICU patients is evident, but using these models is not the only vital factor on the

subject of improving prediction accuracy. The two other essential factors in enhancing prediction accuracy are feature selection and handling imbalanced class distribution problem. The class imbalance distribution is a common problem for medical data, and it can affect model performance [26]. Therefore, due to the importance of these factors, Roumani *et al.* [27] compared the performance of several general data mining methods handling imbalanced data problem. Later, García *et al.* [28] and Liu *et al.* [29] concentrated on dimensionality reduction as well as handling the imbalanced class problem, and they achieved excellent results in mortality prediction. A summarized list of research works on the mortality prediction of intensive care unit patients is presented in detail in Table 1.

There is an apparent lack of use of ensemble and hybrid methods. A combination of several methods and predictive models helps to improve machine learning results. This approach leads to better predictive performance compared to a single model. This study tries to propose a new hybrid model using the Genetic Algorithm with a new ensemble model based on the combination of Stacking and Boosting methods to develop a robust early mortality prediction model while handling the imbalanced data problem.

The unique innovations and significant processes of the present research as compared to similar works include:

- Proposing a new hybrid model based on the Genetic Algorithm and a new ensemble classifier (the combination of Stacking and Boosting methods) where the Genetic Algorithm is used as a feature selection method.
- Comparing the proposed model with the different single and ensemble machine learning models.
- Applying feature scaling to standardize the range of independent features of data.
- Handling the imbalanced data as one of the significant problems in the field of machine learning using the SVM-SMOTE method.
- Applying both Random Hold-Out and Shuffle 5-Fold Cross-Validation methods to perform the validation step.
- Measuring the performance of the implemented models using different evaluation methods, including Accuracy, Area Under the ROC Curve, Recall, Precision, and F1-Score.
- Validating the results by analyzing the differences between all classifiers and indicating the best classifier among others using the Friedman Test as a statistical significance test.
- Comparing the performance of the new proposed hybrid model with different scoring systems such as APACHE II and SAPS II.
- Comparing the performance of the new proposed hybrid model with the state-of-the-art predictive models applied to the MIMIC III dataset as a benchmark dataset for experimental validation.

The rest of this paper is formed as follows: The next section explains the dataset and all the utilized preprocessing

TABLE 1. A summarized list of research works on the mortality prediction of intensive care unit patients.

Article	Machine Learning Model						Ensemble Model			Heuristic Search Algorithm	Hybrid Method	Imbalance Data Problem	Validation		Statistical Evaluation	Feature Selection
	Logistic Regression	K-Nearest-Neighbor	Naive Bayes	Support Vector Machine	Artificial Neural Network	Decision Tree	Bagging	Stacking	Boosting	Genetic Algorithms			Hold-Out	K-Fold Cross		
Doig, Inman [17]	✓				✓								✓			
Dybowski, Gant [16]	✓				✓					✓			✓			
Clermont, Angus [18]	✓				✓								✓			
Silva, Cortez [19]	✓				✓								✓			
Luaces, Taboada [21]	✓			✓											✓	
Kim, Kim [23]	✓			✓	✓								✓			
Xia, Daley [15]					✓							✓		✓		✓
Roumani, May [27]	✓			✓		✓						✓	✓			
Moridani, Setarehdan [20]				✓	✓								✓			
Ghose, Mitra [24]							✓								✓	
Houthoof, Ruyssinck [22]		✓		✓	✓	✓	✓						✓			✓
Awad, Bader-El-Den [13]			✓			✓	✓						✓		✓	
Darabi, Tsinis [25]					✓										✓	
Liu, Chen [29]	✓			✓	✓		✓						✓		✓	
Present Work	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

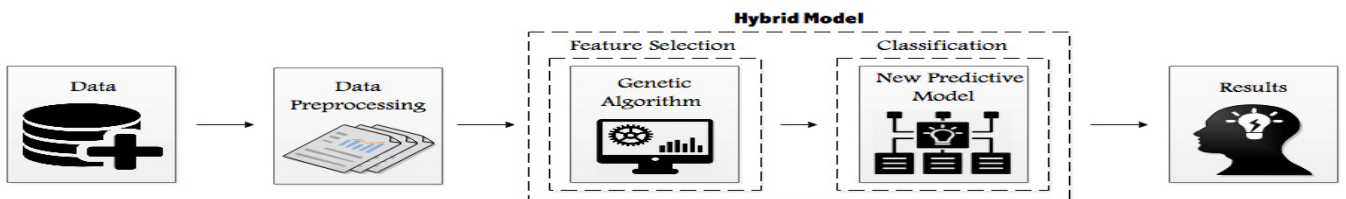


FIGURE 1. Different steps of the proposed methodology.

methods, such as handling the imbalanced data and feature scaling. Section 3 presents the details about the new hybrid machine learning model. In section 4, evaluation methods are described as a way to analyze the performance of the classifiers. Section 5 gives the results and comprehensive analysis to explain the performance of the proposed hybrid model compared to other predictive models. Finally, Section 6 reveals the conclusion and recommends some directions for future research.

II. MATERIAL & METHODS

A key part of the research is understanding the problem. This paper tries to develop a new hybrid model to predict mortality risk in ICU patients in the first stage of their arrival in ICU. It should be noted that all submitted models are coded in

Python, which is an interpreted, high-level, general-purpose programming language. Furthermore, all practical experiments are carried out with a 2 GHz Intel Core i7 MacBook Pro with 4GB of RAM. The applied methodology to achieve the purpose of this paper is depicted in Figure 1.

A. DATASET INFORMATION

The selected data for this research is collected and recorded manually from the hospitals related to Shahid Beheshti University of Medical Sciences and Health Services in Iran between 2013 and 2019. This dataset is the first information collected after the arrival of the patient in ICU (First 24 hours). Therefore, the features are considered constant throughout the whole study. This dataset contains 1999 records and 21 attributes. The number of 20 attributes

TABLE 2. Main features of the studied Intensive Care Unit dataset.

Feature Name	Description	Type
Age	The age of the patients in years	Numeric
Sex	The gender of the patients (0 for Men and 1 for Women)	Nominal
Body Temperature	The body temperature of the patients (Normal Range is 36.5–37.5 °C)	Numeric
Systolic Blood Pressure	The pressure in the blood vessels when the heartbeats	Numeric
Diastolic Blood Pressure	The pressure in the blood vessels when the heart rests between beats	Numeric
Pulse Rate	The number of times a person's heartbeats per minute (Heart Rate)	Numeric
Respiratory Rate	The number of breaths a person takes per minute	Numeric
HCO ₃ (Bicarbonate)	The byproduct of the body metabolism that helps in regulating pH, or Acid balance	Numeric
Blood PH	A logarithmic scale used to specify the acidity or basicity of blood	Numeric
Na (Sodium)	Blood Sodium that helps maintain healthy blood pressure and regulates whole fluid balance	Numeric
K (Potassium)	Blood Potassium that balances the effects of sodium and helps keep fluid levels within a certain range	Numeric
Cr (Creatinine)	Blood creatinine reveals essential information about how well the kidneys are working	Numeric
Hct (Hematocrit)	The volume percentage of red blood cells in the blood	Numeric
WBCs (White Blood Cells)	The number of white blood cells that help fight infections	Numeric
GCS (Glasgow Coma Scale)	The most well-known scoring system used to express the level of awareness in a person	Numeric
PTT (Partial Thromboplastin Time)	A blood test that characterizes coagulation of the blood (Results measured in seconds)	Numeric
Ventilation	Representation of using a ventilator in the ICU to assist in patients breathing (0 for Not Using and 1 for Using)	Nominal
Addiction	Representation of Drug addiction in patients (0 for Not Addicted and 1 for Addicted)	Nominal
Postoperative Patient	Represent the situation of the ICU patient whether if the patient is under postoperative care or not (0 for No and 1 for Yes)	Nominal
Diabetes	Represent the situation of the ICU patient whether if the patient has diabetes or not (0 for No and 1 for Yes)	Nominal
Mortality Status	Represent the final status of the ICU patient whether if the patient is alive or not (0 for Dead and 1 for Alive)	Nominal

¹ Source of Data: ICUs of hospitals related to Shahid Beheshti University of Medical Sciences and Health Services in Iran (2013-2019)

are related in mortality prediction, and one attribute serves as an output or the predicted variable. Table 2 details the main features of the dataset.

B. DATA PREPROCESSING

Data preprocessing is an essential step in machine learning and data mining. The quality of data affects the learning ability of a machine learning model. Therefore, it is so important to prepare the data before feeding it to the model [30]. Data preprocessing is a technique that is used to change the raw data into a clean dataset [31]. It should be noted that the introduced dataset has no missing data, so handling the missing points as a step of data preprocessing is not needed.

1) IMBALANCED DATA PROBLEM

One of the main barriers to machine learning is the imbalanced data problem. This problem happens when the classes are not represented equally [32]. In the event of imbalanced

data, the majority classes dominate the minority classes. On account of this fact, the machine learning classifiers are more biased towards majority classes, and the machine learning models are much more likely to classify new observations to the majority class. As a result, the imbalanced data problem can cause poor classification for minority classes [33], [34].

It should be pointed out that the introduced ICU dataset is significantly imbalanced, and it includes more samples from one class (1517 cases of survival) while the other class is much smaller (only 482 cases of death). Accordingly, the classifier may perform too defective to get results, and it is essential to handle the imbalanced problem.

A variety of techniques have been developed to solve the imbalanced data problem that can be executed during the preprocessing step. Resampling is one of the most generally used approaches to increase the number of minority instances and create a new dataset [35]. This method includes under-sampling and over-sampling

techniques. Over-sampling confirms and implies the minority class by creating new samples or repeating the old ones, while under-sampling removes the samples from the majority class [36].

One of the commonly over-sampling methods, which helps to solve the imbalanced data problem, is SMOTE method [26]. This method generates new samples by interpolating based on the distances between the point and its nearest neighbors. The SMOTE method determines the distances for the minority samples near the decision boundary and creates new samples, so the decision boundary is induced to move further away from the majority classes and prevent the overfitting problem [37], [38]. This paper overcomes the imbalanced data problem using an over-sampling technique named SVM-SMOTE. This method is known as one of the best over-sampling techniques, which has shown better performance than most of the resampling methods. This method really helps the predictive models to present excellent and trustable performance [39], [40]. In 2020, Ghorbani and Ghousi [41] compared various resampling techniques such as Borderline SMOTE, Random Over Sampler, SMOTE, SMOTE-ENN, SVM-SMOTE, and SMOTE-Tomek to handle the imbalanced data problem while using different datasets. The results of their research work reveal that the SVM-SMOTE is more efficient than the other resampling methods, and this method improves the performance of classifiers. The SVM-SMOTE generates new minority class instances near borderlines with the help of SVM to set the boundary between classes and notice to data distribution and density information, which is essential to synthesize minority classes.

Cross-validation is a model validation technique applied to assess how the statistical analysis results are generalized into an independent dataset [42].

This paper uses two general forms of cross-validation, which are random hold-out and shuffle 5-fold cross-validation. The hold-out method randomly divides the 80% of data into the training set and 20% of data into the test set. Also, shuffle 5-fold cross-validation randomly splits the dataset into five equal-sized subsets, uses one of the subsets as the test set, and the other four subsets as the training set and repeats this hold-out strategy five times. It is a well-established fact that the imbalanced class should only be fixed on the training set, and the test set classes should not be touched at all. Therefore, the SVM-SMOTE is only applied to the training set while using hold-out and shuffle 5-fold cross-validation.

2) FEATURE SCALING

Feature scaling is a way of standardizing the range of independent variables, which is also known as data normalization. Most commonly, the datasets contain highly diverse features in sizes, units, and range. Since most of the machine learning models use Euclidean distance, it can affect the performance of the models [43], [44]. The range of ICU dataset points used in this paper is widely varied; therefore, feature scaling

is necessary to suppress the mentioned effects on the performance of models. This paper uses standardization as a method to perform feature scaling.

In this method, the features are rescaled; consequently, all of them have the characteristics of a standard normal distribution with $\mu = 0$ and $\sigma = 1$ where μ is the average, and σ is the standard deviation from the average. The standard scores of the samples are measured as follows [45]:

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

III. MACHINE LEARNING MODELS

There are several kinds of machine learning models to solve classification problems [46]. Various machine learning models including Random Forest [47], K-nearest-neighbor [48], Artificial Neural Network [49], [50], XG-boost [51], [52], Support Vector Machine (Polynomial, Linear, Radial Basis Function, and Sigmoid kernels) [53], [54], Decision Tree [55], [56], Logistic Regression [57], and Naïve Bayes [58] are carried out after the preprocessing step to be compared with a new proposed hybrid model. All of the machine learning models used in this paper are listed in Table 3, together with their best parameters' settings.

TABLE 3. Machine learning models with their specific parameters' settings.

Methods	Parameters
Artificial Neural Network	2 Hidden Layer, Activation Function = rectified linear unit, Maximum Iterations = 200
K-Nearest Neighbor	N_Neighbors = 7, Weight Function = distance, Leaf_Size = 30
Random Forest	N_Estimators = 200, Min_Samples_Leaf = 1
Logistic Regression	C = 13, Penalty = l2, Solver = liblinear
Decision Tree	Criterion = entropy, Splitter = best
Naïve Bayes	No Parameters
Support Vector Machine	C=2.5, Kernel = rbf, Gamma= scale
XG-Boost	N_Estimators = 68, Loss Function = deviance, Learning_Rate = 1.0

A. NEW HYBRID MODEL

This paper introduces a new hybrid model while using the Genetic Algorithm as a feature selection method and a new ensemble model based on the combination of Stacking and Boosting ensemble methods. Feature Selection is one of the fundamental concepts which significantly influences the performance of the model in machine learning [59]. This process selects an optimal subset of relevant features to be used in the development of predictive models. Mainly, feature selection techniques can reduce the dimensionality of the dataset by ignoring the insignificant or noisy features so that

the predictive models can be more accurate. Feature Selection methods are sub-categorized into Filter and Wrapper methods. The wrapper methods select an optimal feature subset using the classifier while the filter methods select the features without using the classifier. This paper uses The Genetic Algorithm based Wrapper Feature Selection (GAWFS) to determine the optimal subset of features. Actually, the selection of GA as a feature selection method in this paper is because of the results of the various research works that have compared different feature selection methods. Most of these research works have mentioned that GA is an excellent feature selection method [60], [62], [63]. The Genetic Algorithm is a heuristic optimization technique inspired by the procedures of natural evolution. Genetic Algorithm evaluates each individual's fitness of a population that is a set of individuals and chromosomes (a subset of features). Particular features are selected based on a fitness function. The fitter individuals have more chance to be kept in the next generation or be chosen for the recombination pool. This paper uses the initial number of 100 for the population and 20 for the generation. The AUC metric is used in GAWFS for calculating the fitness value associated with a particular feature subset. AUC does not bias on size of test or evaluation data, so it can be considered as a better measure of classifier performance than accuracy. AUC measures the overall quality of a classifier. The process of applied Genetic Algorithm based Wrapper Feature Selection is illustrated in Algorithm 1.

Algorithm 1 Process of Genetic Algorithm Based Wrapper Feature Selection

Input: Set of all features related to the ICU mortality

Output: Optimal subset of features

1. **While** iterations ≤ 20 **do** (Number of Generations = 20)
 2. Generate a feature set randomly (Number of population = 100 and Each feature subset represents an individual chromosome)
 3. Select parents and implement genetic operations
 4. Create a new generation
 5. Evaluate the fitness of new generation (Using evaluation measure methods)
 6. **If** the performance of new-fitness $>$ old-fitness **Then** new generation's fitness is better
 7. Replace the current generation with the new generation
 8. **Else**
 9. Keep the current generation
 10. **End If**
 11. **End While**
-

Every machine learning model is intended to better estimate the output variable. The prediction error for different machine learning models can be divided into Bias error and Variance error. Bias is how far are the predicted values from the actual values. On the other hand, the variance occurs when the model performs well on the trained dataset but does not do well on a dataset that is not trained on. Achieving

low bias and low variance is the key to excellent prediction performance, but increasing the bias will decrease the variance and increasing the variance will decrease bias. Finding a balance between bias and variance is needed to minimize the total error and get a good prediction. The idea of ensemble methods is to attempt reducing bias and variance of classifiers by combining several classifiers to create a robust model that obtains better performance [64].

Stacking is an ensemble learning technique, which combines information from multiple predictive models to develop a new model. In contrast to a single model, this approach offers better predictive performance. It is worth mentioning that the combining mechanism in the Stacking is that the output of the base classifiers (Level 0) will be used as training data for another classifier named Meta-Classifier (Level 1) to approximate the same target function. The aim of Stacking is to ensemble strong, diverse sets of classifiers together [65]. Boosting is a constant process that each model attempts to correct the errors of the previous model. Therefore, the following models are dependent on the previous model. Consequently, the Boosting algorithm combines some weak classifiers to develop a robust classifier. The original models would not perform well on the entire dataset, but they work great for some parts of the dataset. Thus, each model boosts the performance of the ensemble model [66].

It should be noted that selecting a base model is needed to use the Boosting algorithm. This algorithm uses a base classifier with a different distribution to find a weak rule [52]. In fact, the base learner derives a random weight distribution (W) for training examples to make the wrongly-classified samples more critical. The model trains other classifiers based on this weight distribution, and the weight distribution will be adjusted again and again. Each time base learner generates a new weak prediction rule. Therefore, after many iterations, the Boosting algorithm combines these weak rules into a single powerful prediction rule.

There are different combinations of machine learning models to build a Stacking ensemble model. This paper has examined the various combinations of models, one by one, to achieve the best prediction result. After testing different combinations, the Multilayer Perceptron Neural Network (MLP), K-Nearest-Neighbor, Extra Tree Classifier are chosen as base classifiers. Moreover, this paper applies adaptive boosting while using the number of 57 Support Vector Machine (RBF Kernel) as the base classifier. The Boosted Support Vector Machine (RBF Kernel) is selected as the meta-classifier. Therefore, this paper links three simple classifiers with another ensemble model to create a new powerful ensemble model. The procedure of the Stacking-Boosting ensemble model is shown in Algorithm 2.

IV. EVALUATION METHODS

Evaluating the machine learning algorithm is an indispensable part of implementing the predictive models. There are several kinds of performance measures to choose from. This paper uses Accuracy (Since Error Rate is equal to one minus

Algorithm 2 Process of New Stacking-Boosting Ensemble Model

Input: Training Data $D = \{x_i, y_i\}_{i=1}^m$ ($x_i \in \mathbb{R}^n$)
Output: A new ensemble classifier H

1. **Step 1:** Learn first level classifiers
2. **For** $t \leftarrow 1$ to T **do** (T : Number of classifiers in the first level)
 3. Learn the base classifier h_t based on D
4. **End For**
5. **Step 2:** Construct new dataset from $D \rightarrow (D')$
6. **For** $i \leftarrow 1$ to m **do** (m : Number of records in the dataset)
 7. Construct a new dataset that includes $\{x'_i, y_i\}$ while $x'_i = \{h_1(x_i), h_2(x_i), \dots, h_T(x_i)\}$
8. **End For**
9. **Step 3:** Learn second level classifier
10. Learn a new classifier h' using Boosting method based on the newly constructed dataset (D')
11. Initialize the weight distribution W
12. **For** $c \leftarrow 1$ to C **do** (C : Number of classifiers in the second level)
 13. Learn weak classifier h_c based on D' and W_c
 14. Evaluate weak classifier $\varepsilon(h_c)$
 15. Update weight distribution W_{c+1} based on $\varepsilon(h_c)$
16. **End For**
17. **Return** $H(x) = h'(\{h_1(x_i), h_2(x_i), \dots, h_T(x_i)\})$

Accuracy, the error rate can be calculated too), Area Under the ROC curve (AUC), Recall, Precision, and F1-Score as metric systems of measurement. Moreover, statistical significance testing is applied to examine the differences between classifiers.

A. STATISTICAL EVALUATION

Comparing machine learning models is a critical operation. Using evaluation measures is simple, but the results can be misleading. The challenge with selecting the best model is to determine how much the estimated capabilities of each model can be trusted. Statistical significance tests are planned to address this problem [67]. The repeated-measures ANOVA is the general statistical test method to analyze the differences between more than two related sample means. The null hypothesis in this test is that all classifiers perform the same, and the detected differences are hardly random [68]. ANOVA is based on three assumptions, but in analyzing the performance of machine learning models, ANOVA assumptions are most probably violated. These assumptions are as follows:

- 1- The drawn samples should be normally distributed.
- 2- The sample cases should be independent of each other.
- 3- The variance between the classifiers should be approximately equal.

This paper uses the Kolmogorov-Smirnov normality test [69] to assess the normality of data. This test examines the Empirical Cumulative Distribution Function (ECDF) of data

with the distribution expected if the data were normal. The null hypothesis of this test is that the data follow a normal distribution; therefore, if the p-value of this test is less than α ($\alpha = 0.05$), the null hypothesis will be rejected, and the data is not normal.

It is well known that ANOVA assumptions may be violated. In this case, the Friedman test, which is a non-parametric equivalent of the repeated-measures ANOVA, can be used [70]. This test is used to analyze the differences between classifiers. The null-hypothesis being examined in this test is that all classifiers perform the same, and rejection of the null hypothesis indicates that one or more of the paired classifiers has a different performance. This paper uses the accuracy data obtained by shuffle 5-fold cross-validation for each classifier. The Friedman test procedure ranks the data of each fold together, then considers the values of ranks by classifiers [71]. Therefore, this test gives a sum of ranks for each classifier that helps to determine the most effective classifier, among others.

V. RESULTS & DISCUSSION**A. MACHINE LEARNING MODEL RESULTS**

There are many different machine learning models; nevertheless, this paper used the most well-known and prominent machine learning models to compare with the new hybrid model. The newly introduced model is applied separately with and without using GA as a feature selection method to give a better perspective of the performance of the proposed model and the impact of the feature selection. It should be noted that only the Hybrid Model (GA + new model) uses the optimal subset of the feature since GA is a part of this model as a feature selection method. Model validation applied in this paper is based on the random hold-out and shuffle 5-fold cross strategies.

1) HOLD-OUT METHOD RESULTS

As mentioned, this paper splits 80% of data into the training set and 20% of data into the test set using the hold-out method. Furthermore, SVM-SMOTE is applied into the training set to handle the imbalanced data problem. Table 4 indicates the performance and running time of the different machine learning models and the newly designed model using the hold-out strategy. Also, Figure 2 shows the test accuracy results for a better perception of the difference among the performances.

Accuracy is the most common evaluation method to measure the performance of a classifier. This metric system of measurement is easy to understand. However, it disregards many vital factors that should be considered in assessing the performance of a model, and it is not enough to fairly judge the model. Table 4 results show that the new hybrid model performs well with test set accuracy. The combination of GA and new ensemble model has significantly improved the accuracy compared to other classifiers. The new model is the only classifier that could achieve accuracy higher than 80%.

TABLE 4. Performance of the models based on the 80/20 random hold-out strategy.

Model	Test Set Accuracy	AUC	Recall	Precision	F1-Score	Running Time
					Weighted Average	
GA + New Model (Hybrid Model)	82.50%	76.33%	98%	86.12%	88.68%	~ 3090 Seconds
New Model	79.00%	69.78%	86.79%	85.87%	79.00%	~ 210 Seconds
Random Forest (RF)	77.25%	68.83%	85.66%	84.26%	77.05%	≤ 1 Second
Artificial Neural Network (ANN)	77.00%	69.66%	84.33%	84.89%	77.07%	~ 27 Seconds
Support Vector Machine (SVM-RBF Kernel)	76.50%	68.00%	85.00%	83.88%	76.33%	≤ 1 Second
XG-Boost	73.50%	63.33%	83.66%	81.49%	73.12%	≤ 1 Second
K-Nearest-Neighbor (KNN)	72.50%	62.00%	83.00%	80.84%	72.10%	≤ 1 Second
Logistic Regression (LR)	67.75%	67.49%	68.00%	86.00%	69.72%	≤ 1 Second
Decision Tree (DT)	67.50%	60.33%	74.66%	80.57%	68.49%	≤ 1 Second
Naïve Bayes (NB)	66.25%	67.49%	65.00%	86.66%	68.44%	≤ 1 Second

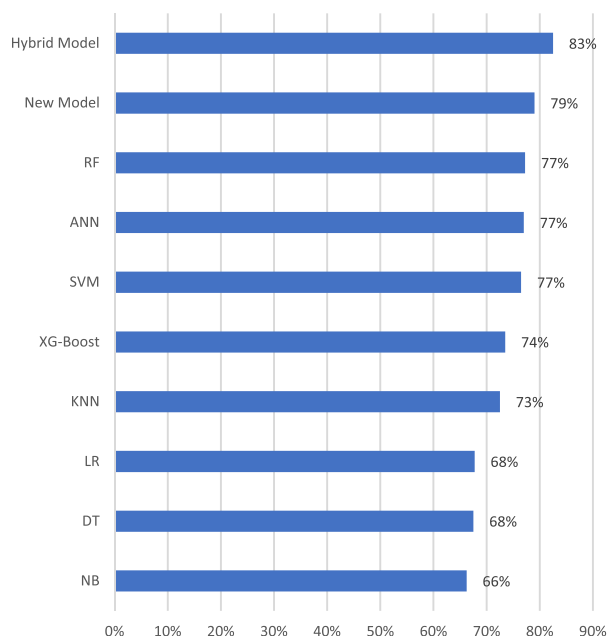


FIGURE 2. The test accuracy results models based on the 80/20 random hold-out strategy.

Also, the performance of the new model using all the features (without applying GA) is better than other machine learning models. The comparison between the results of the hybrid model and the proposed model indicates the impact of the GA as a feature selection method. It seems that by using the novel hybrid model, the accuracy could be improved by 3.50%, which is excellent. Since this paper handled the imbalanced data problem, the accuracy can be more reliable. However, it is better not to consider accuracy as a base measure of performance in this problem.

Accuracy deals with ones and zeros (the model either predicted the class label right or didn't). But many predictive

models can quantify their uncertainty about the answer by outputting a probability value. The model needs to consider a threshold to decide when zero turns into one for computing accuracy from probabilities. AUC is a vital evaluation metric for assessing the performance of classification models. This metric considers all possible thresholds. AUC reveals how much a model is proficient at distinguishing between the different classes. The higher the AUC, the better the model's performance at distinguishing between patients, so the new hybrid model with 76.33% of the AUC metric result has the best performance among the other classifiers. This result means that the new model can differentiate between survival and dead patients with 76.33% of chance, which is significantly better than other classifiers. Furthermore, after the hybrid model, the new model (without applying GA) with 69.78% of the AUC metric shows better performance than other predictive models. It appears that by using the novel hybrid model, the AUC is increased almost more than 7%.

Precision and Recall are useful ways to assess prediction efficiency. The Recall explains the completeness of the tests, while Precision shows how beneficial the outcomes are. The Precision attempts to answer the following question: "What proportion of positive identifications was actually correct?". Moreover, Recall tries to answer the following question: "What proportion of actual positives was identified correctly?". It should be noted that to evaluate the effectiveness of a model sufficiently, you must consider both Precision and Recall. Unfortunately, Precision and Recall are often in tension, and improving Precision typically reduces Recall [72], [73].

The new hybrid model has the best results of Recall and Precision among all other models. The results show that the new hybrid model achieved 98% with the Recall test and 86.12% with Precision; therefore, the hybrid model

TABLE 5. Performance of the models based on the shuffle 5-fold cross-validation.

Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average Accuracy	Variance
GA + New Model (Hybrid Model)	80.75 %	83.25 %	81.50 %	79.50 %	81.63 %	81.32 %	1.00 %
New Model	82.50%	81.25%	78.75%	74.50%	77.94%	78.98%	2.00 %
Random Forest	78.00%	78.75%	77.00%	78.75%	74.43%	77.38 %	1.00 %
XG-Boost	74.75%	79.75%	77.25%	76.00%	75.43%	76.63%	1.00 %
Support Vector Machine (RBF Kernel)	76.25%	75.25%	74.75%	74.50%	71.17%	74.38 %	1.00 %
Artificial Neural Network	75.00%	74.75%	72.25%	71.25%	71.92%	73.03 %	1.00 %
K-Nearest-Neighbor	71.75%	71.00%	71.75%	70.75%	67.91%	70.63 %	1.00 %
Logistic Regression	65.50%	70.25%	66.00%	62.50%	60.43%	64.73 %	3.00 %
Decision Tree	65.75%	63.50%	67.25%	61.75%	59.50%	63.55%	3.00 %
Naïve Bayes	66.50%	63.00%	64.25%	58.25%	54.19%	61.23 %	4.00 %

correctly identifies 98% of all survival patients in ICU, and when it predicts the status of a patient as survival, it is correct 86.12% of the time. Moreover, the new model (without applying GA) outperforms other models with the Precision and Recall metrics too. Analyzing and comparing the models by the Recall and the Precision is delicate, so using the F1-score method is a way to solve this problem. F1-score is the Precision and the Recall harmonic average taking account of both metrics, which determines how accurate and authoritative the model performs. The weighted average of F1-score confirms that the new ensemble model with 88.68% has the best F1-score compared to other classifiers. These results emphasize the excellent performance of the new proposed hybrid model with and without applying GA.

The optimal feature subset selected by GA in the process of using the hybrid model (using random hold-out) consists of 7 features, including Ventilation, Postoperative Patient, Diabetes, Sex, Pulse Rate, Respiratory Rate, HCO₃ (Bicarbonate). Therefore, the new ensemble model has considered these seven features to reach a better prediction. It seems that these features are related to each other. Using a ventilator in the ICU helps patients breathe, which is important, and the doctors of the hospitals have confirmed it. The results indicate that the Age of the patient is not so significant in predicting the mortality risk. Also, Systolic Blood Pressure, Diastolic Blood Pressure, Blood PH, Na (Sodium), K (Potassium), Cr (Creatinine), Hct (Hematocrit), WBCs (White Blood Cells), GCS (Glasgow Coma Scale), PTT (Partial Thromboplastin Time), and addiction can be ignored in predicting mortality. All of the information about these features are explained in detail in table 1.

2) K-FOLD CROSS VALIDATION METHOD RESULTS

Another method of validation is k-fold cross-validation. This research uses shuffle 5-fold cross-validation, which divides the dataset into five subsets. Table 5 shows the obtained results implementing machine learning models with shuffle 5-fold cross-validation. It displays the achieved accuracy of each fold by different models. Moreover, the achieved average accuracy and variance are also presented and compared.

The results of shuffle 5-fold cross-validation point out that the proposed hybrid model obtained the best results among the other models. The lowest accuracy achieved by the proposed model is related to the fourth fold, 79.50%, and the highest accuracy is associated with the second fold, which is 83.25%. The hybrid model has reached an average accuracy of 81.32% with a low amount of 1.00% variance; therefore, the accuracy results are exceptional, and the new model's performance is excellent and acceptable. Also, the new model (without using GA) has shown high performance while using shuffle 5-fold cross-validation. The new model has achieved an average accuracy of 78.98% with a low amount of 2.00% variance, which is the best performance after the hybrid model. These results prove that the new proposed model has an excellent prediction performance, and using GA as a feature selection method to create a hybrid model improves the results notably. Also, Figure 3 displays the comparison among different implemented machine learning models while using shuffle 5-fold cross-validation and hold-out.

The optimal feature subset selected by GA in the process of using a hybrid model (using shuffle 5-fold cross-validation) consists of 10 features, including Age, Ventilation, Sex, Postoperative Patient, Diabetes, Na (Sodium), Hct (Hematocrit), WBCs (White Blood Cells), HCO₃ (Bicarbonate), and PTT

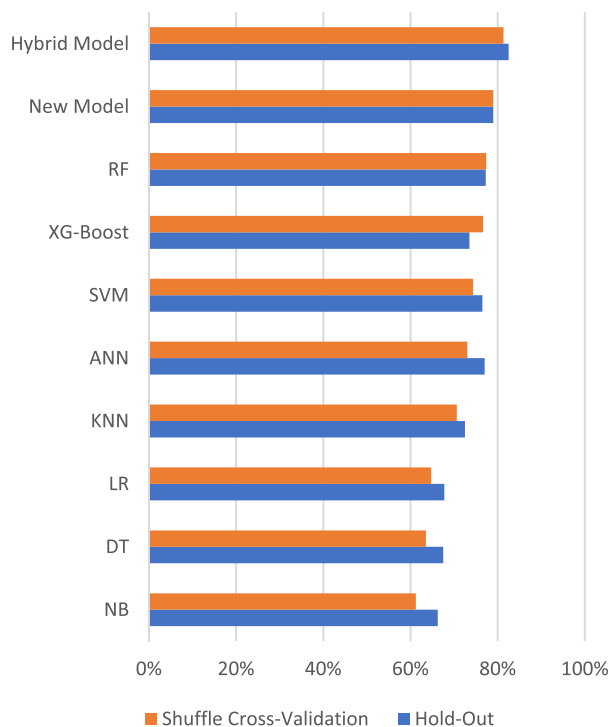


FIGURE 3. Comparing shuffle 5-fold cross-validation and hold-out accuracy results.

(Partial Thromboplastin Time). Therefore, the new ensemble model has considered these ten features to reach a better prediction. It should be regarded that the doctors of the hospitals, where the data is collected, believed that using a ventilator, having diabetes, and being under postoperative care are among the essential features by their experience in different medical situations. Considering the optimal subsets chosen while using random hold-out and shuffle 5-fold cross-validation, it seems that Ventilation, Sex, Postoperative Patient, Diabetes, and HCO₃ (Bicarbonate) are among the most important features because they are selected in both validation methods, so the results can be regarded reliable.

3) STATISTICAL TEST RESULTS

Statistical significance tests are planned to handle the challenge of selecting the best model. As stated, this paper uses the accuracy data collected by shuffle 5-fold cross-validation for each classifier. Some assumptions should be met before using the ANOVA test. First, ANOVA assumes that the samples are drawn from normal distributions. The Kolmogorov-Smirnov normality test results determine that the p-value is 0.049, which is less than 0.050 ($\alpha = 0.050$); therefore, the null hypothesis is rejected, and the ANOVA assumption is not met. Table 6 represents the results of the Kolmogorov-Smirnov normality test.

Due to the Kolmogorov-Smirnov normality test results, ANOVA normality assumption is violated; accordingly, ANOVA does not seem to be a suitable statistical test for this machine learning study. Therefore, the Friedman test is

TABLE 6. The Kolmogorov-Smirnov normality test results.

	Mean	Standard Deviation	Number of samples	P-value
Tested Data	71.42%	7.04 %	45	0.049

TABLE 7. The Friedman test results.

Friedman test	Degrees of freedom	Chi-Square	P-Value
	8	37.55	0.000

used to compare machine learning classifiers. Table 7 shows the results of the Friedman test. These results indicate that the p-value is 0.000. Because the p-value for the classifiers' accuracy data is less than the significance level of 0.05, the null hypothesis is rejected, which means that at least one of the classifiers has a different performance.

Table 8 shows the results of the median and sum of ranks obtained by the Friedman test. The median is the midpoint of the dataset. This midpoint value is the point where half of the data points are above the value, and half of the data points are below the value. Moreover, the overall median is the median of all data points. The median response for the new hybrid model is substantially higher than the overall median. Furthermore, the result of the sum of ranks for the new model is better than other classifiers. These results confirm that the new hybrid model might be more effective than the different classifiers.

TABLE 8. Additional information from Friedman test results.

Rank	Classifiers	Median	Sum of Ranks
1	New Hybrid Model	81.51	45.0
2	Random Forest	77.67	37.0
3	XG-Boost	77.40	36.0
4	Support Vector Machine (RBF Kernel)	74.75	30.0
5	Artificial Neural Network	73.42	27.0
6	K-Nearest-Neighbor	70.92	20.0
7	Logistic Regression	64.81	12.0
8	Decision Tree	63.58	11.0
9	Naïve Bayes	61.51	7.0
	Overall	71.73	

B. COMPARING DIFFERENT SCORE SYSTEMS WITH NEW HYBRID MODEL

This paper compares the performance of the newly introduced hybrid model with different scoring systems such as Acute Physiology and Chronic Health Evaluation II

TABLE 9. The APACHE II and SAPS II performance results.

Scoring Systems	Overall Performance	
	Brier Score(min-max) - [STD]	AUC
APACHE II	0.17 (0-0.94) - [0.25]	74.50 %
SAPS II	0.196 (0-0.999) - [0.35]	75.10 %

(APACHE II) [9] and Simplified Acute Physiology Score II (SAPS II) [10]. After calculating the APACHE II and SAPS II scores, the Brier score (overall performance) and Area under the Receiver Operating Characteristic Curve (AUC) are displayed in table 9. The results reveal that APACHE II is associated with better overall performance. It should be pointed out that the lower the Brier score is, the better the performance will be. Moreover, it seems that the SAPS II system performs better with AUC. As mentioned, the new proposed hybrid model has achieved 76.33% with AUC measure, which is better than both of these scoring systems.

C. COMPARISON ON A BENCHMARK DATASET

This paper uses the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC) III database, a real-world health care dataset, as a Benchmark dataset to test the new hybrid model [74]. This dataset consists of 58976 patients records and 17 features related to ICU patients. There are many non-numeric variables in this dataset that hide and mask lots of interesting information. So, this paper used some different methods, such as converting to number (converting a categorical into the numerical variables) and dummy coding (converting a categorical input variable into a continuous variable) to deal with such variables. Over the years, several approaches are proposed to predict mortality risk in ICUs. This paper implemented the new proposed hybrid model on the MIMIC III dataset to benchmark the new model’s performance against the current state-of-the-art works in this domain. Research such as Calvert *et al.* [75], Purushotham *et al.* [76], Che *et al.* [77], Darabi *et al.* [25], Harutyunyan *et al.* [78], and Zhang [79] are the state-of-the-art machine learning-based models that were developed and benchmarked MIMIC III dataset. The new hybrid model is applied to the MIMIC III dataset, and its accuracy and AUC performances are compared to the related state-of-the-art works. Table 10 shows the results of this comparison.

The results represent that the proposed hybrid model outperforms all the state-of-the-art models applied to the MIMIC III dataset in terms of both accuracy and AUC measures. This model achieves 98.20% with the accuracy-test and 88.47% with the AUC test, which is better than other state-of-the-art models. Some of the state-of-the-art models have not reported their model’s prediction accuracy, so it was not possible to indicate their accuracy. It can be concluded that

TABLE 10. Comparison of new model performance against state-of-the-art works applied to the MIMIC III dataset.

State-of-the-Art Study	Year	Accuracy		AUC	
		Study	New Model	Study	New Model
Calvert, Mao [75]	2016	80.00 %		88.00 %	
Purushotham, Meng [76]	2017	Not Reported		86.73 %	
Che, Purushotham [77]	2018	Not Reported	98.20 %	84.00 %	88.47 %
Darabi, Tsinis [25]	2018	Not Reported		87.30 %	
Harutyunyan, Khachatrian [78]	2019	Not Reported		87.00 %	
Zhang [79]	2019	Not Reported		82.10 %	

the introduced hybrid model is useful in predicting mortality risk in ICUs.

VI. CONCLUSION

The development of biomedical equipment and healthcare level produces a large amount of data. Finding a way to process this data into useful information can save many lives. Prediction of mortality in the intensive care unit is considered as one of the most vital subjects in healthcare data analysis. An accurate prediction of the mortality risk for ICU patients could provide helpful information about patients’ lives and reduce costs; therefore, it is critical to predicting it in patients as soon as possible.

This study intends to recommend a new hybrid predictive model using the Genetic Algorithm as a feature selection method and a new ensemble model based on the combination of Stacking and Boosting ensemble methods to create an early mortality prediction model while handling the imbalanced data problem using SVM-SMOTE over-sampling technique. The two methods of random hold-out and shuffle 5-fold cross-validation are used to validate machine learning model stability. Furthermore, The Friedman test, as a statistical significance test, is applied to handle the challenge of selecting the best model. After handling the imbalanced data problem, predictive models are implemented using the random hold-out method on balanced data. The evaluation results confirm that the new hybrid model’s performance is acceptable, and it has the best performance among all other classifiers using different evaluation metrics. Also, the results show that the new model’s performance using all the features (without applying GA) is the second-best performance among all other models. The evaluation results of machine learning models implemented on training balanced data using shuffle 5-fold cross-validation method present the same results as the

random hold-out method related to selecting the best model. The results of shuffle 5-fold cross-validation show that the new model performs better than other models. This model achieved excellent accuracy with a low amount of acceptable variance, which is higher than all classifiers. Additionally, the new model using all the features (without applying GA) outperforms other methods too. The comparison between the results of the hybrid model and the proposed model implies that the GA will improve the performance of the model significantly. The GA technique provides the optimal feature subset (7 features while using random hold-out and ten features while using shuffle 5-fold cross-validation) that improves the performance of the model, and it seems that Ventilation, Sex, Postoperative Patient, Diabetes, and HCO₃ (Bicarbonate) are among the most vital features. The Friedman test results prove that the new hybrid model's performance is better than other classifiers. Comparing the new hybrid model with different scoring systems revealed that the new model has a better performance. Moreover, benchmarking against the state-of-the-art models which predict mortality risk in ICU applied to the MIMIC-III dataset also highlighted the excellent performance of the introduced hybrid model, with an AUC and accuracy improvement over the state-of-the-art approaches.

There are many ways to improve this research, and future works can be carried out in the following directions. The Particle Swarm Optimization (PSO) can be used as a meta-heuristic to optimize the model and improve its performance. Additionally, feature creation can be implemented to construct new features from existing data to help model with better prediction.

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, May 2013.
- [2] 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.
- [3] G. D. Magoulas and A. Prentza, "Machine learning in medical applications," in *Advanced Course on Artificial Intelligence*. Berlin, Germany: Springer-Verlag, 1999.
- [4] R. Danger, I. Segura-Bedmar, P. Martínez, and P. Rosso, "A comparison of machine learning techniques for detection of drug target articles," *J. Biomed. Informat.*, vol. 43, no. 6, pp. 902–913, Dec. 2010.
- [5] J. Xu, Y. Zhang, P. Zhang, A. Mahmood, Y. Li, and S. Khattoon, "Data mining on ICU mortality prediction using early temporal data: A survey," *Int. J. Inf. Technol. Decis. Making*, vol. 16, no. 1, pp. 117–159, Jan. 2017.
- [6] R. Davoodi and M. H. Moradi, "Mortality prediction in intensive care units (ICUs) using a deep rule-based fuzzy classifier," *J. Biomed. Informat.*, vol. 79, pp. 48–59, Mar. 2018.
- [7] A. E. Johnson and R. G. Mark, "Real-time mortality prediction in the intensive care unit," in *Proc. AMIA*, 1998, p. 994.
- [8] G. S. Power and D. A. Harrison, "Why try to predict ICU outcomes?" *Current Opinion Crit. Care*, vol. 20, no. 5, pp. 544–549, Oct. 2014.
- [9] J. E. Zimmerman, "Acute Physiology and Chronic Health Evaluation (APACHE) IV: Hospital mortality assessment for today's critically ill patients," *Critical Care Med.*, vol. 34, no. 5, pp. 1297–1310, 2006.
- [10] J.-R. Le Gall and S. F. J. J. Lemeshow Saulnier, "A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study," *Jama*, vol. 270, no. 24, pp. 2957–2963, 1993.
- [11] S. Lemeshow, "Mortality Probability Models (MPM II) based on an international cohort of intensive care unit patients," *Jama*, vol. 270, no. 20, pp. 2478–2486, 1993.
- [12] F. L. Ferreira, D. P. Bota, A. Bross, C. Mélot, and J.-L. Vincent, "Serial evaluation of the SOFA score to predict outcome in critically ill patients," *J. Amer. Med. Assoc.*, vol. 286, no. 14, pp. 1754–1758, 2001.
- [13] 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. Informat.*, vol. 108, pp. 185–195, Dec. 2017.
- [14] I. Silva, "Predicting in-hospital mortality of icu patients: The physician/computing in cardiology challenge," in *Proc. Comput. Cardiol.*, Sep. 2012, pp. 245–248.
- [15] H. Silva, "A neural network model for mortality prediction," in *Proc. ICU*, vol. 39, 2012, pp. 261–264.
- [16] R. Dybowski, V. Gant, P. Weller, and R. Chang, "Prediction of outcome in critically ill patients using artificial neural network synthesised by genetic algorithm," *Lancet*, vol. 347, no. 9009, pp. 1146–1150, Apr. 1996.
- [17] G. Doig, "Modeling mortality in the intensive care unit: Comparing the performance of a back-propagation, associative-learning neural network with multivariate logistic regression," in *Proc. Annu. Symp. Comput. Appl. Med. Care*, 1993, pp. 1–4.
- [18] G. Clermont, D. C. Angus, S. M. DiRusso, M. Griffin, and W. T. Linde-Zwirble, "Predicting hospital mortality for patients in the intensive care unit: A comparison of artificial neural networks with logistic regression models," *Crit. Care Med.*, vol. 29, no. 2, pp. 291–296, Feb. 2001.
- [19] Á. Silva, P. Cortez, M. F. Santos, L. Gomes, and J. Neves, "Mortality assessment in intensive care units via adverse events using artificial neural networks," *Artif. Intell. Med.*, vol. 36, no. 3, pp. 223–234, Mar. 2006.
- [20] M. K. Moridani, "New algorithm of mortality risk prediction for cardiovascular patients admitted in intensive care unit," *Int. J. Clin. Exp. Med.*, vol. 8, no. 6, p. 8916, 2015.
- [21] O. Luaces, "Predicting the probability of survival in intensive care unit patients from a small number of variables and training examples," *Artif. Intell. Med.*, vol. 45, no. 1, pp. 63–76, 2009.
- [22] R. Houthoofd, J. Ruyssinck, J. van der Herten, S. Stijven, I. Couckuyt, B. Gadeyne, F. Ongenaes, K. Colpaert, J. Decruyenaere, T. Dhaene, and F. De Turck, "Predictive modelling of survival and length of stay in critically ill patients using sequential organ failure scores," *Artif. Intell. Med.*, vol. 63, no. 3, pp. 191–207, Mar. 2015.
- [23] S. Kim and I. R. Kim, "Park, A comparison of intensive care unit mortality prediction models through the use of data mining techniques," *Healthcare Inform. Res.*, vol. 17, no. 4, pp. 232–243, 2011.
- [24] S. Ghose, "An improved patient-specific mortality risk prediction in ICU in a random forest classification framework," *Stud. Health Technol. Inform.*, vol. 214, pp. 56–61, Aug. 2015.
- [25] 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, Apr. 2018.
- [26] S. Fotouhi, S. Asadi, and M. W. Kattan, "A comprehensive data level analysis for cancer diagnosis on imbalanced data," *J. Biomed. Informat.*, vol. 90, Feb. 2019, Art. no. 103089.
- [27] Y. F. Roumani, J. H. May, D. P. Strum, and L. G. Vargas, "Classifying highly imbalanced ICU data," *Health Care Manage. Sci.*, vol. 16, no. 2, pp. 119–128, Jun. 2013.
- [28] M. N. M. García, "Machine learning methods for mortality prediction of polytraumatized patients in intensive care units-dealing with imbalanced and high-dimensional data," in *Proc. Int. Conf. Intell. Data Eng. Autom. Learn.*, 2014, pp. 309–317.
- [29] J. Liu, X. X. Chen, L. Fang, J. X. Li, T. Yang, Q. Zhan, K. Tong, and Z. Fang, "Mortality prediction based on imbalanced high-dimensional ICU big data," *Comput. Ind.*, vol. 98, pp. 218–225, Jun. 2018.
- [30] S. B. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas, "Data preprocessing for supervised learning," *Int. J. Comput. Sci.*, vol. 1, no. 2, pp. 111–117, 2006.
- [31] S. A. Alasadi and W. S. Bhaya, "Review of data preprocessing techniques in data mining," *J. Eng. Appl. Sci.*, vol. 12, no. 16, pp. 4102–4107, 2017.
- [32] X. Guo, "On the class imbalance problem," in *Proc. 4th Int. Conf.*, Apr. 2008, pp. 1–7.
- [33] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Handling imbalanced datasets: A review," *GESTS Int. Trans. Comput. Sci. Eng.*, vol. 30, no. 1, pp. 25–36, 2006.
- [34] W. Hadi, N. El-Khalili, M. AlNashashibi, G. Issa, and A. A. AlBanna, "Application of data mining algorithms for improving stress prediction of automobile drivers: A case study in Jordan," *Comput. Biol. Med.*, vol. 114, Nov. 2019, Art. no. 103474.

- [35] S. Cateni, V. Colla, and M. Vannucci, "A method for resampling imbalanced datasets in binary classification tasks for real-world problems," *Neurocomputing*, vol. 135, pp. 32–41, Jul. 2014.
- [36] 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.*, vol. 291, pp. 184–203, Jan. 2015.
- [37] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.
- [38] S. Maldonado, J. López, and C. Vairetti, "An alternative SMOTE oversampling strategy for high-dimensional datasets," *Appl. Soft Comput.*, vol. 76, pp. 380–389, Mar. 2019.
- [39] Y. Liu and A. X. An Huang, "Boosting prediction accuracy on imbalanced datasets with SVM ensembles," in *Proc. Pacific-Asia Conf. Knowl. Discovery Data Mining*. 2006, pp. 107–118.
- [40] C. Zhang, Y. Zhou, J. Guo, G. Wang, and X. Wang, "Research on classification method of high-dimensional class-imbalanced datasets based on SVM," *Int. J. Mach. Learn. Cybern.*, vol. 10, no. 7, pp. 1765–1778, Jul. 2019.
- [41] 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.
- [42] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. IJCAI*, Montreal, QC, Canada, 1995, pp. 1–7.
- [43] 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, Apr. 2001.
- [44] M. Ebrahimi, M. Mohammadi-Dehcheshmeh, E. Ebrahimie, 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, Nov. 2019, Art. no. 103456.
- [45] K. K. Pal and K. S. Sudeep, "Preprocessing for image classification by convolutional neural networks," in *Proc. IEEE Int. Conf. Recent Trends Electron., Inf. Commun. Technol. (RTEICT)*, May 2016, pp. 1778–1781.
- [46] C. Robert, *Machine Learning, A Probabilistic Perspective*. Cape Town, Tokyo: Taylor & Francis, 2014.
- [47] 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, Mar. 2006.
- [48] P. Cunningham and S. J. Delany, "K-Nearest neighbour classifiers," *Multiple Classifier Syst.*, vol. 34, pp. 1–17, Mar. 2007.
- [49] A. K. Jain, J. Mao, and K. M. Mohiuddin, "Artificial neural networks: A tutorial," *Computer*, vol. 29, no. 3, pp. 31–44, Mar. 1996.
- [50] I. A. Basheer and M. M. Hajmeer, "Artificial neural networks: Fundamentals, computing," *Des. Appl.*, vol. 43, no. 1, pp. 3–31, 2000.
- [51] M. Ziäba, S. K. Tomczak, and J. M. Tomczak, "Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction," *Expert Syst. Appl.*, vol. 58, pp. 93–101, Oct. 2016.
- [52] E. Bauer and R. J. M. L. Kohavi, "An empirical comparison of voting classification algorithms: Bagging, boosting, and variants," *Mach. Learn.*, vol. 36, nos. 1–2, pp. 105–139, 1999.
- [53] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," Dept. Comput. Sci. Inf. Eng., Univ. National Taiwan, Taiwan, Taipei, Tech. Rep., 2003, pp. 1–12.
- [54] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, nos. 1–3, pp. 389–422, 2002.
- [55] S. R. Safavian, "A survey of decision tree classifier Methodology," *IEEE Trans. Syst., Man, Cybern.*, vol. 21, no. 3, pp. 660–674, May/Jun. 1991.
- [56] W. Du and Z. Zhan, "Building decision tree classifier on private data," in *Proc. Int. Conf. Privacy, Secur. Data Mining*, vol. 14, 2002, pp. 1–7.
- [57] D. R. Cox, "The regression analysis of binary sequences," *J. Roy. Stat. Soc.*, vol. 20, no. 1, pp. 215–242, Jul. 1958.
- [58] T. R. Patil, "Performance analysis of Naive Bayes and J48 classification algorithm for data classification," *Int. J. Intell. Syst. Appl. Eng.*, vol. 6, no. 2, pp. 256–261, 2013.
- [59] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Mar. 2003.
- [60] F. E. Fassnacht, C. Neumann, M. Forster, H. Buddenbaum, A. Ghosh, A. Clasen, P. K. Joshi, and B. Koch, "Comparison of feature reduction algorithms for classifying tree species with hyperspectral data on three central European test sites," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2547–2561, Jun. 2014.
- [61] B. Xue, M. Zhang, and W. N. Browne, "A comprehensive comparison on evolutionary feature selection approaches to classification," *Int. J. Comput. Intell. Appl.*, vol. 14, no. 02, Jun. 2015, Art. no. 1550008.
- [62] F. Tan, X. Fu, Y. Zhang, and A. G. Bourgeois, "A genetic algorithm-based method for feature subset selection," *Soft Comput.*, vol. 12, no. 2, pp. 111–120, Sep. 2007.
- [63] T. Nyathi and N. Pillay, "Comparison of a genetic algorithm to grammatical evolution for automated design of genetic programming classification algorithms," *Expert Syst. Appl.*, vol. 104, pp. 213–234, Aug. 2018.
- [64] P. Yang, Y. Hwa Yang, B. B. Zhou, and A. Y. Zomaya, "A review of ensemble methods in bioinformatics," *Current Bioinf.*, vol. 5, no. 4, pp. 296–308, Dec. 2010.
- [65] D. H. Wolpert, "Stacked generalization," *Neural Netw.*, vol. 5, no. 2, pp. 241–259, Jan. 1992.
- [66] Y. Freund, R. Schapire, and N. Abe, "A short introduction to boosting," *J.-Jpn. Soc. Artif. Intell.*, vol. 14, nos. 771–780, p. 1612, 1999.
- [67] J. Demár, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, Jan. 2006.
- [68] R. A. Fisher, "Statistical methods and scientific inference," in *The Founders of Evolutionary Genetics* (Boston Studies in the Philosophy of Science), vol. 142, S. Sarkar, Eds. Dordrecht, The Netherlands: Springer, 1956.
- [69] R. H. W. Lilliefors, "On the kolmogorov-smirnov test for normality with mean and variance unknown," *J. Amer. Stat. Assoc.*, vol. 62, no. 318, pp. 399–402, Jun. 1967.
- [70] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *J. Amer. Stat. Assoc.*, vol. 32, no. 200, pp. 675–701, Dec. 1937.
- [71] M. Friedman, "A comparison of alternative tests of significance for the problem of m rankings," *Ann. Math. Statist.*, vol. 11, no. 1, pp. 86–92, Mar. 1940.
- [72] C. W. Cleverdon, "On the inverse relationship of recall and precision," *J. Documentation*, vol. 28, no. 3, pp. 195–201, Mar. 1972.
- [73] M. Buckland and F. Gey, "The relationship between recall and precision," *J. Amer. Soc. for Inf. Sci.*, vol. 45, no. 1, pp. 12–19, Jan. 1994.
- [74] A. E. W. Johnson, T. J. Pollard, L. Shen, L.-W.-H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database," *Sci. Data*, vol. 3, no. 1, Dec. 2016, Art. no. 160035.
- [75] J. Calvert, Q. Mao, J. L. Hoffman, M. Jay, T. Desautels, H. Mohamadlou, U. Chettipally, and R. Das, "Using electronic health record collected clinical variables to predict medical intensive care unit mortality," *Ann. Med. Surgery*, vol. 11, pp. 52–57, Nov. 2016.
- [76] S. Purushotham, C. Meng, Z. Che, and Y. Liu, "Benchmark of deep learning models on large healthcare MIMIC datasets," 2017, *arXiv:1710.08531*. [Online]. Available: <http://arxiv.org/abs/1710.08531>
- [77] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," *Sci. Rep.*, vol. 8, no. 1, p. 6085, Dec. 2018.
- [78] H. Harutyunyan, H. Khachatrian, D. C. Kale, G. Ver Steeg, and A. Galstyan, "Multitask learning and benchmarking with clinical time series data," *Sci. Data*, vol. 6, no. 1, p. 96, Dec. 2019.
- [79] Z. Zhang, "Prediction model for patients with acute respiratory distress syndrome: Use of a genetic algorithm to develop a neural network model," *PeerJ*, vol. 7, p. e7719, Sep. 2019.



RAMIN GHORBANI received the M.S. degree in system optimization (data mining in healthcare) from the Iran University of Science and Technology, Iran, in 2019. He is currently pursuing the Ph.D. degree with the Pattern Recognition and Bioinformatics Group, Delft University of Technology (TU Delft). His research interests include artificial intelligence, machine learning, and data analysis, with a passion for health informatics and bioinformatics.



tems and services, human factor engineering, health, safety, and environment management system, time, and work-study.

ROUZBEH GHOUSI received the Ph.D. degree from the Iran University of Science and Technology, in 2013. He has been a Faculty Member at SIE, since 2015. He is currently an Assistant Professor of industrial engineering with the Iran University of Science and Technology. His research interests include safety and healthcare engineering, human reliability, sustainable supply chain management, and data science. He teaches human reliability analysis, diagnosis of production systems and services, human factor engineering, health, safety, and environment management system, time, and work-study.



ALIREZA ATASHI is the Head of the Cancer Informatics Department, Motamed Cancer Institute, Tehran, Iran, and an Assistant Professor in medical informatics at the e-Health Department, Tehran University of Medical Sciences. His research interests include medical data mining, clinical informatics, patient registries, and telemedicine, especially in cancer and intensive care fields in which he has done several studies.

...



serves as the Editor-in-Chief for the *Journal of Industrial and Systems Engineering* (JISE) and *Decision Science Letters*.

AHMAD MAKUI received the Ph.D. degree from the Iran University of Science and Technology, in 2000. He has been a Faculty Member at SIE, since 2003. He is currently a Professor of industrial engineering with the Iran University of Science and Technology. His research interest includes operations research and its applications, especially linear programming. He teaches operations research, decision making analysis, and production planning courses. Besides, he currently