Kütahya Dumlupınar University
Institute of Graduate Studies

Journal of Scientific
Reports-A
E-ISSN: 2687-6167
Number 54, September 2023

RESEARCH ARTICLE

CHARACTERIZATION OF MORTALITY PREDICTION: AN ENSEMBLE LEARNING ANALYSIS USING THE MIMIC-III DATASET

Anıl Burcu ÖZYURT SERİM[1,*]

[1,*] İstanbul Haliç University, Management Information Systems, İstanbul, Türkiye, burcuozyurt@halic.edu.tr,
ORCID: 0000-0001-98682676

Receive Date: 23.08.2023 *Accepted Date: 30.09.2023*

ABSTRACT

Applications that employ medical data are directly impacted by the classification of imbalanced data. It is vital due to the nature of classification and solutions about medical data. The purpose of this article is to identify a machine learning model that may be successfully applied in the medical field to reduce the number of mortality and optimize the efficiency of hospital resources. For this reason, it is thought that the better the performance of the ML model, the more a different perspective will be gained on the problems in today's medicine. Therefore, in the study, Weighted Random Forest (WRF) and Balanced Random Forest (BRF) which are ensemble machine learning (ML) methods for imbalanced data were implemented to identify the performance of the algorithms for mortality determination from open-source MIMIC-III dataset by using vital signs, comorbidities, and laboratory variables with demographic characteristic information. To evaluate the performance of WRF and BRF, a Random Forest Classifier (RFC) was also implemented to investigate the power of developed models for imbalanced data. In addition, the features used in the ML methods were separated into three groups to explore the impact of the vital signs, comorbidities, and laboratory variables with demographic characteristics separately on mortality identification. In addition to previous applications on UCI datasets, the present study revealed that the BRF method for imbalanced medical data provides high performance in determining the majority and minority classes of the data by using vital signs and laboratory variables with demographic characteristics.

Keywords: *MIMIC-III, Random Forest, Weighted Random Forest, Balanced Random Forest, Ensemble Learning.*

1. INTRODUCTION

Many different data analysis techniques can now be used on massive amounts of data as a result of recent technical advancements. Applying these ideas to the critical care unit is a much more crucial issue because of the intensive care unit's data-rich nature. Decision support systems have made it possible to observe a well-defined set of criteria for rare diseases or unexplained diseases with thorough medical imaging data when they are integrated into ordinary clinical functions [1]. Selection, analysis, and diagnostic interpretation of radiological imaging have been done using the knowledge acquired in this manner. As a result, healthcare personnel are able to treat more patients securely, operate more productively, and make fewer mistakes than ever before because of data-based decision support systems, recently [1].

Also, artificial intelligence and the integration of automated information systems have greatly enhanced medical practices. The open-source MIMIC-II, MIMIC-III, and MIMIC-IV dataset has recently been the subject of studies examining its performance in forecasting hospital mortality in intensive care patients and comparing its performance with that of different algorithms. It is necessary to estimate mortality among intensive care unit (ICU) inpatients in order to assess the severity of a patient's condition and weigh the advantages of cutting-edge therapies, interventions, and healthcare initiatives [2]. In a number of these research, the ability to forecast hospital mortality in intensive care patients using nonparametric methods based on artificial neural networks was assessed. [3, 4]. These studies came to the conclusion that nonparametric methods may be at least as good at predicting ICU mortality as traditional logistic regression [108]. In a related other study, Karun et al. employed the sandwich regression method to determine the characteristics that increase the risk of pneumonia in patients receiving intensive care. The study found that BMI, kidney disease, hypertension, diabetes, and asthma are some of the major risk factors for pneumonia in the elderly. According to the Poisson regression findings, he also discovered that the middle-aged group had a larger probability of acquiring pneumonia in the elderly [5]. Pirracchio's [6] study examined if an ensemble machine learning technique called Super Learner would improve hospital mortality prediction for critically ill intensive care patients using data from the Medical Information Mart for Intensive Care II (MIMIC-II). The prediction score generated based on Super Learner was demonstrated to provide better results in terms of both discrimination and calibration when compared with mortality scoring outcomes such as SAPS II (Simplified Acute Physiology Score), APACHE II (Acute Physiologic Assessment and Chronic Health Evaluation), and SOFA (Sequential Organ Failure Assessment) [6, 7]. Support Vector Machine (SVM), Logistic Regression (LR), and XGBoost classification algorithms were employed in the study by Ergul Aydn & Kamişli Ozturk, [8] one of the comparable studies in this field, to assess whether the patients' stays in critical care were longer than 3 days. The study found that, similarly to earlier prediction articles, the XGBoost classifier outperformed Support Vector Machine (SVM) and Logistic Regression (LR) [9, 10, 11, 12]. Poucke's study was to compare the predictive performance of Decision Tree, Naïve Bayes, Logistic and Regression methods, and ensemble learning methods (Random Forest, Boosting, and Bagging) when assessing the predictive power of laboratory tests for hospital mortality in patients admitted to the intensive care unit. In this study using ensemble methods, Random Forests provided the best prediction accuracy for mortality risk prediction, consistent with previous research [13, 14]. According to a study by Yang et al., c-med GAN (conditional medical generative adversarial network) offers a strong classification for predicting mortality in critical care patients when compared to the SAPS II score, SVM (support vector machine), and MLP (multilayer perceptron). When the dataset size was reduced, C-med GAN outperformed MLP in terms of death prediction [2]. Xia et al.'s study, which discovered that over 40% of elderly patients were hospitalized to the intensive care unit, observed that longer hospital stays were associated with higher short- and long-term risks of death [15]. In order to predict ICU mortality, Dybowski et al. and Kim et al. [16, 17] show that nonparametric methods may operate better than conventional logistic regression models.

An important alternative for diagnosing and beginning treatments in the healthcare sector is based on supervised machine learning, recently [18]. Ensemble learning, which takes into consideration themes and attributes, is one of the finest supervised machine learning (ML) approaches to categorizing data. The categorization can be difficult by using basic classification algorithms for the complex data structure including many features [18]. At this point, implementation of innovative techniques such as Balanced Random Forest (BRF) [19] and Weighted Random Forest (WRF) [19] can be a solution since these advanced ensemble learning methods has specific strategies for the datasets that are unbalanced ensuring that each class is given the appropriate level of consideration during model training. Unbalanced data, also known as imbalanced data, means the situation in which the number of instances for each class in the dataset is unequal. In other words, a class has significantly more samples than another class or classes.

In this study, BRF and WRF models, which are the improved versions of one of the successful ensemble model Random Forest Classifier (RFC), were implemented to investigate the feature importance for mortality prediction by using MIMIC-III dataset [20] that can be considered a new approach for identification of the mortality by random forest-based algorithms for imbalanced medicine data. With the presented study authorities can plan preparedness, and allocation of financial resources based on how long patients remain in intensive care units with the help of random forest-based machine learning models. Additionally, it offers suggestions for how to allocate resources effectively, enhance the caliber of healthcare services, and prioritize macro policies in the health sectors of nations by using machine learning methods. Since in medical science, the datasets may have imbalanced structure and limited size, the presented work can be an indicator for the usage of an improved version of random forest-based algorithms for unbalanced data by also considering the feature selection effect on model performances. For this purpose, the features in the data were separated into three groups such as comorbidities, laboratory variables with vital signs, and both to understand the impact of the corresponding properties on mortality of the patients with demographic characteristics. To compare the model performances RFC was also applied to the dataset in the similar way.

2. LİTERATURE REVİEW

In today's evolving technology, the significance of artificial intelligence-based machine learning (ML) approaches is on the rise across diverse domains, including natural sciences [19,21], anomaly detection [22,23], healthcare [24,25], object recognition [26,27], and business [18] since it has a huge impact by radically changing the way of solving sophisticated problems. As a result of the advancement of machine learning algorithms in the field of health, modeling methodologies have increased in variety [24,25]. Predictive models are useful tools to comprehend the underlying causes of diseases and to expand clinical knowledge, the collection and analysis of massive amounts of critical care data is crucial. Large-scale Critical Care databases are valuable tools for learning about individuals' risk factors, regular critical illness history, and the efficacy of various treatment approaches. Because, patients in the critical care unit are more likely than other patients to experience many problems today, and mortality is also higher [28]. Innovations in disease therapeutics and survival rates are essential in this setting because cancer and comorbidity are more closely associated with aging populations in industrialized countries. Various data sources, including clinical records and laboratory findings, are used to get this data [29]. Incorrect or missing data during the data collection phase, although unsynchronized time-referenced data, difficulty processing different data formats, and restrictions on digital storage capacity are the difficulties encountered, the MIMIC-III database, which is easily accessible by the researcher, provides the advantage of not having an access fee, unlike other databases [20,30].

All adult patients admitted to Beth Israel Deaconess Medical Center's intensive care units between 2001 and 2007 are included in this database's first version, MIMIC-II. Clinical data and physiological data make up MIMIC-II's two main parts. Physiological data are digital data produced by electronic device signals recorded for vital signs during the patient's stay in intensive care. Clinical data in the database are organized to include data such as the patient's demographic information, intravenous drug administration rates, and laboratory test results [31]. The MIMIC III dataset includes Beth Israel Deaconess Hospital intensive care unit patients from 2001 to 2012 [20]. The extensive MIMIC III database contains details on individuals older than 16 who are admitted to intensive care units. The data comprises vital signs, medications, lab findings, observations and remarks made by medical personnel, fluid balance, procedure and diagnosis codes, imaging reports, length of hospital stay, survival rates, and demographic information. For specialists doing various studies on intensive care research, from the creation of clinical decision support algorithms to a better comprehension of retrospective clinical investigations, the MIMIC-II and MIMIC-III databases are crucial sources of data.

Despite advances in disease identification and treatment, the rate of mechanical ventilation, sepsis infection and mortality in intensive care units have been growing recently [12,29,32,33]. Globally, deaths in intensive care units are seen as a severe health concern. The onset of disease symptoms in intensive care units of patients at high risk of death, diagnosis with methods with low predictive accuracy, and time-consuming access to laboratory data cause the risk of death to reach its highest level. For this reason, various machine learning models have been developed using data obtained from physicians' risk indicators in intensive care patients. Therefore, earlier disease detection and prediction lead to quicker recovery and better results [34].

3. METHOD

Ensemble learning [18] is one of the most effective techniques for a successful ML model. In ensemble learning, multiple learner objects are created and trained to solve the problem. Basic learners can be decision trees from training data. Ensemble learning, which consists of decision trees, can be performed in the form of bagging. In bagging ensemble learning [35], the ML model is trained with a randomly selected subset of the data. In this approach, the classification is made by voting the outputs of the basic learners. RFC, RFC-based WRF and BRF are the bagging ensemble learning-based models that were used in the study.

3.1. Random Forest Classifier (RFC)

One of the decision tree-based bagging models is the RFC model [36]. RFC is made up of decision trees that act as a community, making judgments based on a variety of sub-decisions and defending one another from individual mistakes. Each decision tree has nodes representing the features in the dataset and leaves representing the algorithm's decisions [19]. Nodes are decided by selecting from the dataset the features to be applied to train the current tree. The model makes sure that the attributes to be used in the nodes for each tree are randomly chosen in order to maintain diversity while minimizing the correlation between trees. The quality of the node separation for each attribute can be evaluated using either the entropy gain or the Gini index, but doing so can entail the risk of misclassification [37].

The RFC method consists of two steps: creating the trees and selecting the decision tree. The initial step is to create decision trees using any randomly chosen model component from the training dataset. After receiving votes from the numerous decision trees in the test set, the final choice is determined in the last phase. The RFC algorithm is shown schematically in Figure 1 [38]. As represented in the figure, the dataset is separated into training and test sets. The RFC model generates the decision trees with the instances in the training set by using the features as random nodes which allows the model to learn the classes with patterns. The algorithm evaluates the test data by the constructed decision trees during training. It categorizes test data according to the most votes from the decision trees.
The pseudo code of the algorithm is represented below in which ‘X’, ‘y’ and, ‘n_estimators’ represent the input features, the corresponding target labels, and the number of trees (number of iterations), respectively.

Algorithm 1: Random Forest Classifier

```
Initialize with a number of trees (n_estimators)

Train the model:
for each tree in n_estimators:
    Create a decision tree using some data (X, y)
    Add the tree to the ensemble

Make predictions:
for each tree in the ensemble:
    Get predictions from the tree
Combine the tree predictions to make a final prediction
```

The RFC model has many advantages over other ML techniques. One of the main problems in the application of ML models is that the model, which is defined as overfitting, cannot correctly classify data outside the training set. The RFC method is resistant to over-learning because it has a forest of decision trees. Since the importance of each feature in the decision-making process can be calculated, the model can be interpreted and the relationship between the features used and the decisions can be recognized more clearly. The application of the RFC model has the potential to produce successful results for particle identification and event selection [37,38].

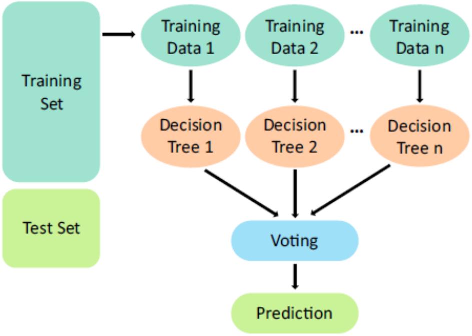


Figure 1. Schematic representation of the RFC classification model [38]. The dataset is divided into training and testing. The algorithm constructs the decision trees with the examples in the training set by using the features in the data as random nodes. It is ensured that the model learns about classes with patterns. Evaluates the completed RFC test dataset according to the created decision trees. It classifies the model test data according to the category with the most votes from the decision trees.

3.2. Weighted Random Forest (WRF)

The WRF model is a class-weighted RFC model adapted to manage groups of different sizes in the dataset. The ARO method was developed to increase the success of the RFC algorithm on skewed datasets [39]. In this model, in order to reduce the effect of majority samples in the dataset on learning and to increase the learning of the minority group, a weight inversely proportional to the size of the classes in the dataset is used in the nodes and predictions of the decision trees in the model [39]. The pseudo code of the WRF is shown in the following where 'X', 'y' and, 'n_estimators' represent the input features, the corresponding target labels, and the number of trees (number of iterations), respectively.

Algorithm 2: Weighted Random Forest

```
Initialize with a number of trees (n_estimators)

  Train the model:
  for each tree in n_estimators:
    Create a decision tree using some data (X, y) and consider sample weights
    Add the tree to the ensemble

  Make predictions:
  for each tree in the ensemble:
    Get predictions from the tree
  Combine the tree predictions to make a final prediction
```

This method has attracted great interest for the classification of data, which often has data imbalances. Unlike the current random forest method, class weights are based on the assignment of separate weights for each class, rather than a single weight. With high accuracy in categorizing majority and minority classes, this recommended approach provides a solution to the problem of classifying between majority and minority classes for unbalanced medical data. It also shows that it improves the overall performance of the classifier.

3.3. Balanced Random Forest (BRF)

The BRF technique is another version of the RFC model developed for skewed datasets. BRF applies a subsampling technique for each decision tree generation process in the RFC algorithm. This is why it is known as the Balanced Random Forest because it combines the sampling technique with the idea of an ensemble where unbalanced data processing becomes an algorithmic process. In the BRF algorithm, the size of the majority group in the dataset is determined randomly and in accordance with the minority group size. The classifier tackles imbalanced data by creating decision trees that are sensitive to both majority and minority classes. For the majority class, it employs balanced sampling, ensuring a representative subset of majority class instances in each tree. For the minority class, it ensures that all minority class instances are included. This approach leads to decision trees that can capture relevant patterns in the minority class while maintaining a fair balance with the majority class, ultimately resulting in a more equitable and accurate classification of both classes when combined in the forest. The pseudo code of the BRF is illustrated below where 'X', 'y' and, 'n_estimators' represent the input features, the corresponding target labels, and the number of trees (number of iterations), respectively.

Algorithm 3: Balanced Random Forest

```
Initialize with a number of trees (n_estimators)

  Train the model:
  for each tree in n_estimators:
    Create a balanced training sample from the data (X, y)
    Create a decision tree using the balanced sample
    Add the tree to the ensemble

  Make predictions:
  for each tree in the ensemble:
    Get predictions from the tree
  Combine the tree predictions to make a final prediction
```

There are studies in the literature comparing the recall or sensitivity of its performance which revealed that BRF performed better than the Random Forest algorithm [40,41]. Although there are studies in which RF's ensemble learning algorithm is applied to unbalanced data for classification [42], it is known that BRF offers a better approach to the classification problem in these unbalanced data with multiple classification problems [40]. The computational efficiency of BRF over WRF for skewed data was also demonstrated by Chen et al. [39].

4. ANALYSİS

In the analysis, MIMIC-III dataset were studied to investigate the impact of the features on mortality prediction of RFC based ML models. In addition, different groups of features were used in each method application to understand the effect of various categorical properties on decisions of the machine learning algorithms.

4.1. Dataset

MIMIC-III, a substantial and openly accessible database, contains health-related data devoid of personal identification for over 40,000 patients who received care within critical care units at the Beth Israel Deaconess Medical Center from 2001 to 2012 [20]. The following information was included as a feature in the analysis by taking into account prior studies [43,44,45,46], clinical relevance, and available data: comorbidities (hypertension, atrial fibrillation, ischemic heart disease, diabetes mellitus, depression, hypoferric anemia, hyperlipidemia, chronic kidney disease (CKD), and chronic obstructive pulmonary disease [COPD]); and laboratory variables (hematocrit, red blood cells, mean corpuscular hemoglobin [MCH], mean corpuscular hemoglobin concentration [MCHC], mean corpuscular volume [MCV], red blood cell distribution width [RDW], platelet count, white blood cells, neutrophils, basophils, lymphocytes, prothrombin time [PT], international normalized ratio [INR], NT-proBNP, creatine kinase, creatinine, blood urea nitrogen [BUN] glucose, potassium, sodium, calcium, chloride, magnesium, the anion gap, bicarbonate, lactate, hydrogen ion concentration [pH], partial pressure of CO2 in arterial blood, and LVEF) [47]. The dataset includes demographic information like age, gender, and body mass index (BMI) at admission, as well as vital indicators including heart rate (HR) systolic blood pressure [SBP], diastolic blood pressure [DBP] respiration rate, body temperature, saturation pulse oxygen [SPO2], and urine output in the first 24 hours [47]. The description of each feature is represented in Table 1 [48].

Table 1. The description of each feature used in the analysis [48].

Name	Description	Unit	Min	Max
age	Patient age	years	0	100
basos	basophils	%	0	50
bicar	bicarbonate	mEq/L	5	50
bmı	body mass index	kg/m^2	18.5	24.9
bun	Blood urea nitrogen	mg/dL	0	200
ca	calcium	mg/dL	4	20
ck	Creatine kinase	IU/L	0	-
cl	chloride	mEq/L	80	130
crea	creatinine	mg/dL	0	15
dbp	Diastolic blood pressure	mmHg	0	200
gender/ sex	patient sex	-	-	-

glu	glucose	mg/dL	0	1000
hct	hematocrit	%	15	60
hgb	hemoglobin	g/dL	4	18
hr	Heart rate	bpm	0	300
ınr	international normalized ratio	prothrombin time/international normalized ratio	-	-
k	potassium	mEq/L	0	10
lact	lactate	Mmol/L	0	50
mg	magnesium	mg/dL	0.5	5
lvef	left ventricular ejection fraction	%	52	72
mch	mean corpuscular hemoglobin	pg	0	-
mchc	mean corpuscular hemoglobin concentration	%	20	50
mcv	mean corpuscular volume	fL	50	150
na	sodium	mEq/L	110	165
neut	neutrophils	%	0	100
NT-proBNP	N-terminal pro–B-type natriuretic peptide	pg/mL	0	100
spo2	Oxygen saturation	%	50	100
pco2	partial pressure of CO2 in arterial blood	mmHg	10	150
pH	hydrogen ion concentration	-	6.8	8
plt	platelet count	K/uL	5	1200
pt	prothrombin time	sec	0	-
rbc	red blood cells	m/uL	0	20
rdw	red blood cell distribution width	%	0	100
resp	respiration rate	insp/min	0	120
sbp	systolic blood pressure	mmHG	0	300
temp	temperature	C	32	42
urine	urine output	mL	0	2000
wbc	white blood cells	K/uL	0	-
hp	hypertension	mmHg	90/60	120/80
A-Fib	atrial fibrillation	dk	60	100
ihd	ischemic heart disease	mmol/L	1.20	1.62
iddm	diabetes mellitus	mg/dL	0	100
bdı	Depression (Beck Depression Inventory)	-	-	-
hwa	hypoferric anemia	ng/mL	11	-
hdl	hyperlipidemia	mg/dL	0	200
ckd	chronic kidney disease (CKD	mL/min	60	-
copd	chronic obstructive pulmonary disease [COPD]	%	88	92

For the analysis, the information of the patients having missing value was removed from the dataset. In the study, 428 records including information of 65 dead (15% of the dataset) and 363 (85% of the dataset) alive patients were analysed with 48 features and outcome information which represents mortality conditions. Table 2 explains the number of instances for each class in total, train and test dataset which indicates that each group has 15% mortality samples.

Table 2. Mortality and alive class sizes for total, train, and test datasets.

Dataset	Total	Train	Test
Alive Class	363	292	71
Mortality Class	65	50	15

4.2. Application of Ensemble Learning

In the analysis, WRF, and BRF algorithms were used to analyze the mortality condition of the patients by using three different feature group. In addition, RFC algorithm was also implemented to compare its performance with its improved versions for imbalanced data. In group 1 all features given in the dataset were utilized by the algorithms as properties to predict the condition of the patient. In group 2 only comorbidities were considered as features by the models. In group 3 laboratory variables with vital signs were used as features by the methods. In all feature groups, demographic characteristics were included. In all models 100 trees (number of iteration) were used [49], and the quality of the node separation was preferred “Gini” for all algorithms. In WRF, the parameter used for setting the class weights was set to ‘balanced’ which modifies the weights inversely proportional to class instances in the input data [50]. Due to the structure of the data 80% of the data was selected for training and the rest for the test.

Figures 2, 3, and 4 represent the significance of features in RFC, WRF, and BRF across three distinct groups. Analyzing the outcomes of models used group 1 features, it is evident that apart from age, vital attributes exert minimal influence on the algorithms. Additionally, the relevance of comorbidity information in group 1 is below 2%. The comparison of feature importance within group 1 leads to the observation that similar attributes wield a shared impact on algorithmic decisions. Plot illustrations concerning group 2 reveal that ‘age’ and ‘bmi’ emerge as dominant features across all algorithms. Interestingly, comorbidity features exerted a modest influence on models, accounting for less than 10%. Analysis of feature importance plots within group 3 suggests that 'urine output' is the pivotal feature in BRF and WRF algorithms, while 'bmi' and 'age' consistently affect decision mechanisms in all models. Notably, gender information held negligible importance in the decision algorithms across all group analyses.

5. RESULTS AND DISCUSSIONS

In the study, sensitivity or recall, precision, and F1 score, performance metrics of the ML, were used for the comparison of model performances by considering the class sizes in which the mortality and alive information were labeled as positive class (signal class) and negative class, respectively. In addition, AUC ROC [51] values were also determined to assess the discrimination power of the algorithms for unbalanced datasets [52,53].

5.1. Performance Metrics

Sensitivity (recall), precision, and F1 score values will be determined using parameters found in the confusion matrix, a table of the real class information in the dataset against the ML model predictions, obtained as a result of the ML model application [54,55]. Recall or sensitivity is used to calculate the percentage of fatalities that may have been accurately predicted, by Equation (1).

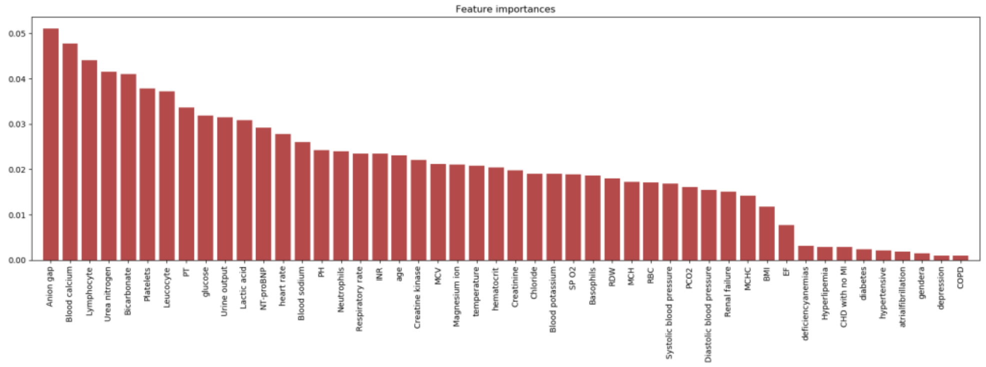
$$\text{Sensitivity} = \text{TP} / (\text{TP} + \text{FN}) \tag{1}$$

In Equation (1), The values for TP and FN correspond to the proportion of information that was mistakenly categorized as alive and correctly classified as mortality, respectively. The expression "precision" is used to describe the accuracy of the predictions made by the model using Equation (2) in which FP represents misclassified mortality samples.

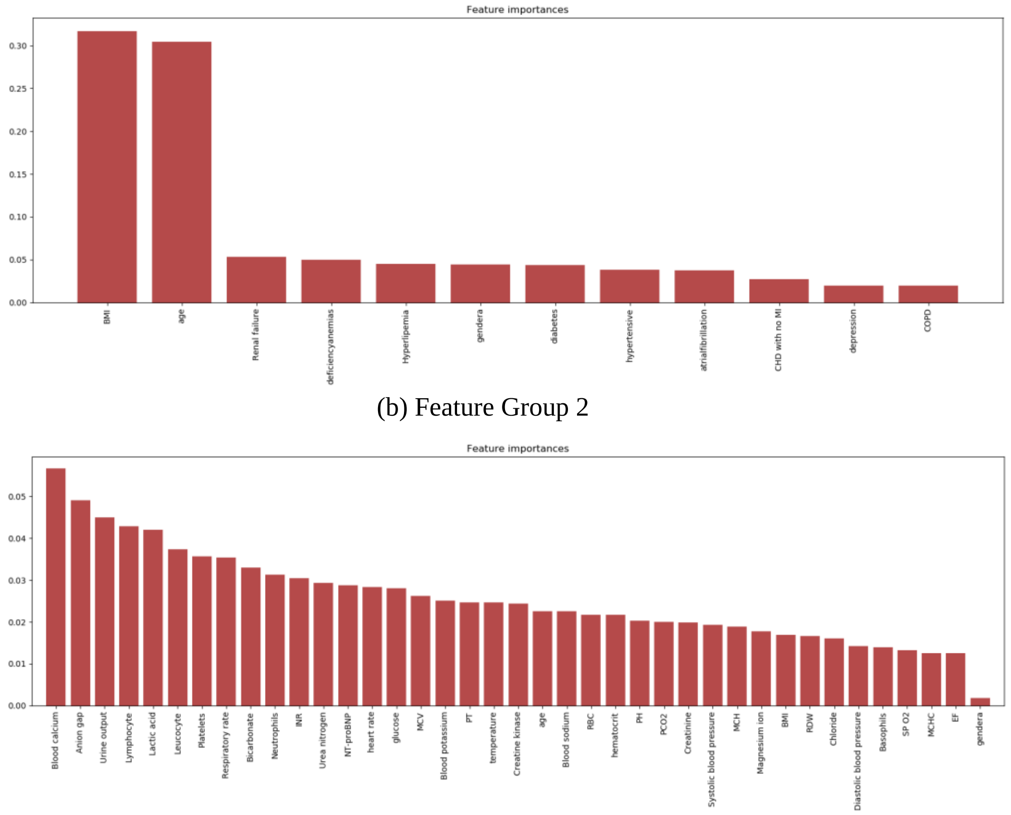
$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \tag{2}$$

The harmonic mean of sensitivity and precision, known as the F1 score, allows for a comprehensive two-sided evaluation of the model represented in Equation (3).

$$\text{F1 score} = (2 * \text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \tag{3}$$



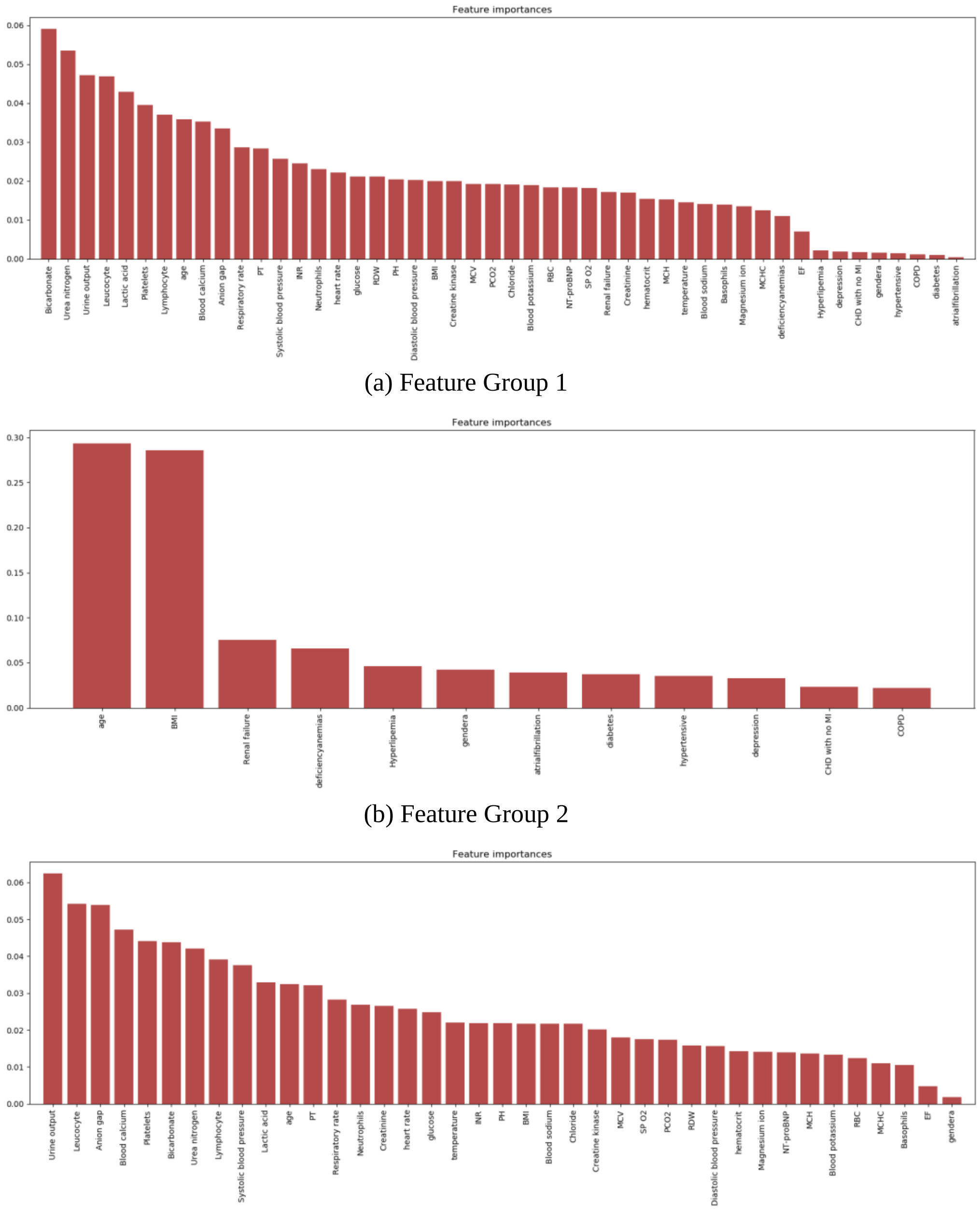
(a) Feature Group 1



(b) Feature Group 2

(c) Feature Group 3

Figure 2. The feature importance in RFC for three feature groups.

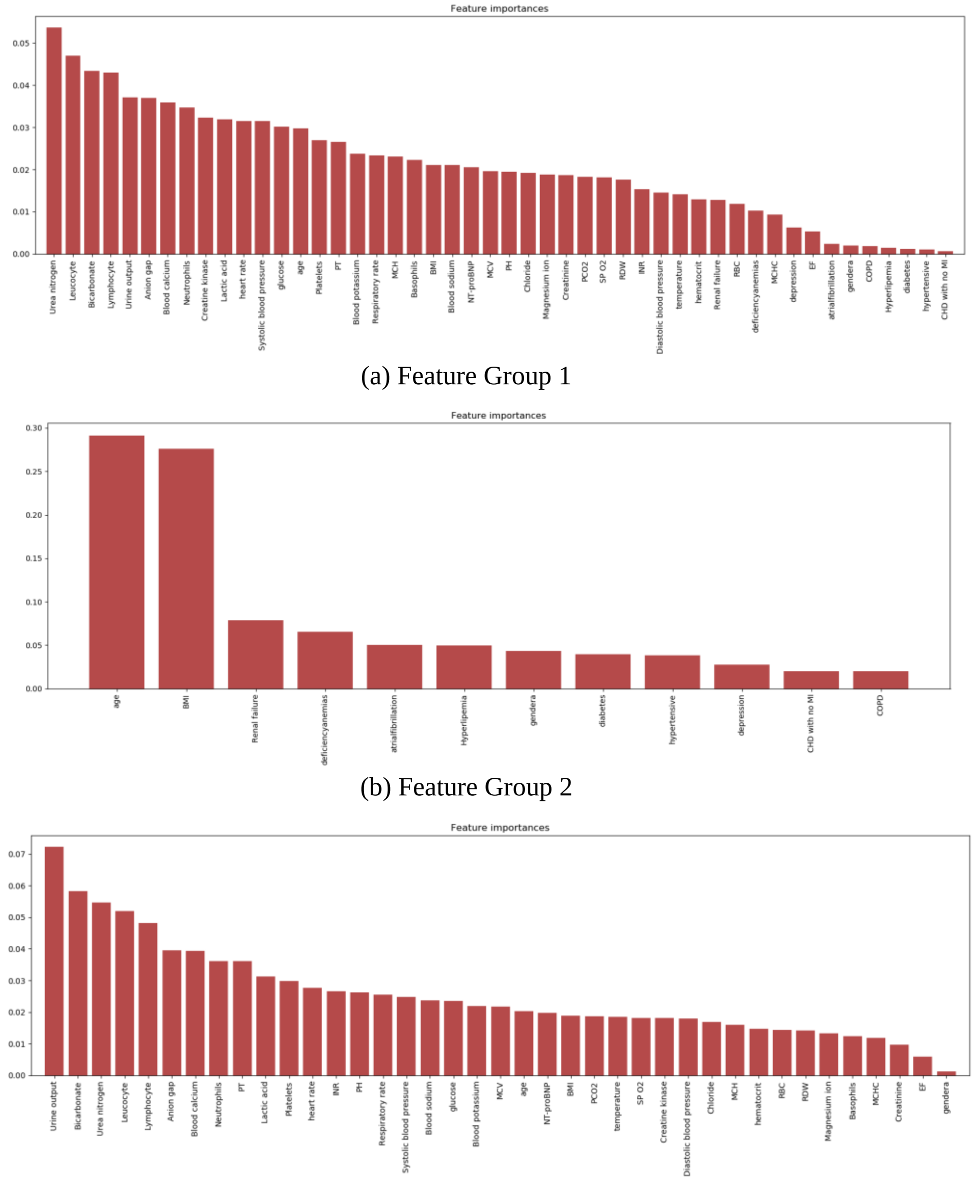


(a) Feature Group 1

(b) Feature Group 2

(c) Feature Group 3

Figure 3. The feature importance in WRF for three feature groups.



(a) Feature Group 1

(b) Feature Group 2

(c) Feature Group 3

Figure 4. The feature importance in BRF for three feature groups.

The Receiver Operating Characteristic Curve (ROC) curve, one of the most important metrics in machine learning, illustrates the relationship between the rate of misclassified positive classes and the rate of mortality instances. The method includes several criteria to identify model performance. The ROC curve is based on determining (iii) two-dimensional graphs of (i) sensitivity or recall and (ii) "false positive rate (FPR=FP/(FP+TN) where TN is correctly categorized alive class.)" criteria and (iv) calculating the area under the curve (AUC ROC). An indicator of how well a metric can discriminate between two diagnostic classes is the area under the ROC curve (AUC ROC). AUC ROC is a measure of how well two classes can be distinguished by the model. A high AUC ROC value signifies an improvement in the model's ability to differentiate between mortality and alive classes [52,53]. If the values of F1 score, and AUC ROC close to 1, that means the model is a perfect classifier [54]. Accuracy, another performance evaluation metric of the ML model, was not used in the analysis since it might not provide sufficient information about the success of models applied to datasets with imbalanced classes [39,40,41].

5.2. Evaluation of Model Performances

Table 2 represents RFC, WRF, and BRF weighted performances by implementing three different features in the algorithms to find the mortality from MIMIC-III dataset with AUC ROC values. The model performances can be comprehended by closeness of each metric to the value 1. For an ideal case, perfect classifier exhibits the maximum performance scores, 1. For performance comparison of machine learning models, the algorithm is outperformed if the scores are higher than the rest. The AUC ROC score implies the success of the algorithm and precision score represents how well the model categorizes the classes. According to the table, 88.095% of BRFs that used group 3 attributes were able to accurately identify mortality at 90.814%. A higher F1 score indicates better model performance which is BRF model using Group 3 features. The table illustrates how all algorithms have low success rates when comorbidities are taken into account as features. The results exhibited that even if RFC is not modified for the dataset its achievement is very close to the improved versions when the discriminative features were used. In general usage of Group 3 features enables all the algorithms to find each class with high precision. The results confirm that BRF handles imbalanced data by training its decision trees in a way that gives equal importance to both the majority and minority classes, ensuring a fair and accurate classification for both.

Table 2. Table of RFC, WRF, and BRF model performances. The performance metrics' top scores are shown in bold.

Model	**Feature Group**	**Precision**	**Recall**	**F-1 score**	**AUC ROC**
RFC	Group 1	0.86773	0.87209	0.84850	0.65962
	Group 2	0.80349	0.83721	0.80894	0.58631
	Group 3	0.88064	0.86047	0.81939	0.60000
WRF	Group 1	0.78904	0.83721	0.78242	0.52877
	Group 2	0.77829	0.82558	0.78940	0.55060
	Group 3	0.89756	**0.88372**	0.84587	0.58333
BRF	Group 1	0.86893	0.73256	0.77716	0.71842
	Group 2	0.77054	0.65116	0.69279	0.59028

	Group 3	**0.90814**	0.84884	**0.86390**	**0.88095**

6. CONCLUSIONS

In recent years, digital health record system data has become a valuable research area for data analysis with ML approach. It is thought that machine learning (ML) approaches will provide solutions to the issues brought on by a lack of evidence in their application domains by improving the accessibility of large-scale medical datasets. In addition, potential patients who can be treated are given the opportunity to better predict events such as heart attacks and death in intensive care using machine learning methods. Thanks to mortality estimates, hospitals can now more accurately anticipate resource needs, properly identify illnesses, and decide whether a patient needs additional care before it's too late. [55,56].

In the present study, compared to RFC and WRF models, the results of BRF algorithm in which the laboratory variables with vital signs were used showed the highest performance for the morality determination. RFC approach in general provides the best prediction performance in previous studies, however, due to the data structure BRF, the improved version of RFC for imbalanced data, demonstrated higher success than it. WRF model, another version of RFC in which the class weights were arranged, did not show better discrimination power compared to the other models. From the results of models using different features, it can be concluded that comorbidities and gender information did not affect the morality determination. Class imbalance in the classification of medical data is a problem that is currently being studied and has to be solved. In the study, an algorithm (BRF) is proposed to be used in handling imbalanced and limited medical data. BRF, an improved version of one of the successful ensemble learning algorithms RFC, addresses class imbalance by constructing decision trees in a way that ensures balanced representation of minority and majority classes in each tree's training subset. Considering the values of BRF assessment metrics such as AUC ROC and F-1 score, the model is promising to be used in unbalanced medical data analysis.

ACKNOWLEDGEMENTS

The author is grateful to the MIMIC-III Clinical Database for making the data platform free to use for researchers and other academics in educational institutions.

REFERANCES

[1] Hanson III, C. W., & Marshall, B. E. (2001). Artificial intelligence applications in the intensive care unit. Critical care medicine, 29(2), 427-435.

[2] Yang, W., Zou, H., Wang, M., Zhang, Q., Li, S., & Liang, H. (2023). Mortality prediction among ICU inpatients based on MIMIC-III database results from the conditional medical generative adversarial network. *Heliyon*, *9*(2).

[3] Dybowski, R., Gant, V., Weller, P., & Chang, R. (1996). Prediction of outcome in critically ill patients using artificial neural network synthesised by genetic algorithm. The Lancet, 347(9009), 1146-1150.

[4] Gortzis, L. G., Sakellaropoulos, F., Ilias, I., Stamoulis, K., & Dimopoulou, I. (2008). Predicting ICU survival: a meta-level approach. BMC health services research, 8, 1-8.

[5] Karun, K. M., Puranik, A., Lintu, M. K., & Deepthy, M. S. (2023). Risk factors of pneumonia among elderly with robust Poisson regression-A study on mimic III data. Biomedicine, 43(02), 696-700.

[6] Data, M. C., & Pirracchio, R. (2016). Mortality prediction in the icu based on mimic-ii results from the super icu learner algorithm (sicula) project. Secondary Analysis of Electronic Health Records, 295-313.

[7] Eya, J., Ejikem, M., Ogamba, C., & Ogamba, C. M. (2022). Admission and Mortality Patterns in Intensive Care Delivery at Enugu State University of Science and Technology Teaching Hospital: A Three-Year Retrospective Study. Cureus, 14(7).

[8] Aydin, Z. E., & Ozturk, Z. K. (2021). Prediction Length of Stay in Intensive Care Unit in the Presence of Missing Data. Artificial Intelligence Theory and Applications, 1(2), 48-53.

[9] Liu, J., Wu, J., Liu, S., Li, M., Hu, K. & Li, K. (2021) Predicting mortality of patients with acute kidney injury in the ICU using XGBoost model. PLoS One, 16(2). doi: 10.1371/journal.pone.0246306

[10] Leung, W. K., Cheung, K. S., Li, B., et al. (2021) Applications of machine learning models in the prediction of gastric cancer risk in patients after Helicobacter pylori eradication. Aliment Pharmacol Ther, 53 (8), 864– 872.

[11] Pang, X., Forrest, C. B., Lê-Scherban, F., Masino, A. J. (2021) Prediction of early childhood obesity with machine learning and electronic health record data. International Journal of Medical Informatics. 150, 104454. https://doi.org/10.1016/j.ijmedinf.2021.104454.

[12] Silahtaroglu, G., & Canbolat, Z. N. (2020). An early prediction and diagnosis of sepsis in intensive care units: An unsupervised machine learning model. Mugla Journal of Science and Technology, 6(1), 32-40.

[13] Poucke, S. V., Zhang, Z., Schmitz, M., Vukicevic, M., Laenen, M. V., Celi, L. A., & Deyne, C. D. (2016). Scalable predictive analysis in critically ill patients using a visual open data analysis platform. PloS one, 11(1), e0145791

[14] Churpek, M. M., Yuen, T. C., Winslow, C., Meltzer, D. O., Kattan, M. W., & Edelson, D. P. (2016). Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. Critical care medicine, 44(2), 368.

[15] Xia, F., Zhang, J., Meng, S., Qiu, H., & Guo, F. (2021). Association of frailty with the risk of mortality and resource utilization in elderly patients in intensive care units: a meta-analysis. Frontiers in Medicine, 8, 637446.

[16] Dybowski, R., Gant, V., Weller, P., & Chang, R. (1996). Prediction of outcome in critically ill patients using artificial neural network synthesised by genetic algorithm. The Lancet, 347(9009), 1146-1150.

[17] Kim, S., Kim, W., & Park, R. W. (2011). A comparison of intensive care unit mortality prediction models through the use of data mining techniques. Healthcare informatics research, 17(4), 232-243.

[18] Lin, W., Wu, Z., Lin, L., Wen, A., & Li, J. (2017). An ensemble random forest algorithm for insurance big data analysis. *Ieee access*, *5*, 16568-16575.

[19] Yalcin Kuzu, S. (2023). Random Forest Based Multiclass Classification Approach for Highly Skewed Particle Data. *Journal of Scientific Computing*, *95*(1), 21.

[20] Johnson, A., Pollard, T., & Mark, R. (2019). MIMIC-III Clinical Database Demo (version 1.4). PhysioNet. https://doi.org/10.13026/C2HM2Q.

[21] Scheunert, G., Heinonen, O., Hardeman, R., Lapicki, A., Gubbins, M., & Bowman, R. M. (2016). A review of high magnetic moment thin films for microscale and nanotechnology applications. Applied Physics Reviews, 3(1).

[22] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM computing surveys (CSUR), 41(3), 1-58.

[23] Salo, F., Injadat, M., Nassif, A. B., Shami, A., & Essex, A. (2018). Data mining techniques in intrusion detection systems: A systematic literature review. IEEE Access, 6, 56046-56058.

[24] Krawczyk, B., Galar, M., Jeleń, Ł., & Herrera, F. (2016). Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy. *Applied Soft Computing*, *38*, 714-726.

[25] Vuttipittayamongkol, P., & Elyan, E. (2020). Improved overlap-based undersampling for imbalanced dataset classification with application to epilepsy and parkinson's disease. *International journal of neural systems*, *30*(08), 2050043.

[26] Elyan, E., Jamieson, L., & Ali-Gombe, A. (2020). Deep learning for symbols detection and classification in engineering drawings. *Neural networks*, *129*, 91-102.

[27] Zhang, X., Zhuang, Y., Wang, W., & Pedrycz, W. (2016). Transfer boosting with synthetic instances for class imbalanced object recognition. *IEEE transactions on cybernetics*, *48*(1), 357-370.

[28] Tabakoglu, N., & Volkan, I. N. A. L. (2021). Evaluation of Basic Parameters for Prediction of ICU Mortality. Journal of Critical and Intensive Care, 12(2), 47.

[29] Altun, G. T., Arslantas, M. K., Dincer, P. C., Arslantas, R., & Kararmazf, A. (2022). Prognostic value of the lactate–albumin difference for predicting in-hospital mortality in critically ill patients with sepsis. Marmara Medical Journal, 35(1), 61-66.

[30] Harutyunyan, H., Khachatrian, H.D., Kale, C., Ver Steeg,G., Galstyan, A.,(2019). "Multitask learning and benchmarking with clinical time series data," Sci. Data, vol. 6, no. 1, p. 96.

[31] Saeed, M., Villarroel, M., Reisner, A. T., Clifford, G., Lehman, L. W., Moody, G., ... & Mark, R. G. (2011). Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): a public-access intensive care unit database. *Critical care medicine*, *39*(5), 952.

[32] Vincent, J. L., Moreno, R., Takala, J., Willatts, S., De Mendonça, A., Bruining, H., ... & Thijs, L. G. (1996). The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure: On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine (see contributors to the project in the appendix).

[33] LaFaro, R. J., Pothula, S., Kubal, K. P., Inchiosa, M. E., Pothula, V. M., Yuan, S. C., ... & Inchiosa Jr, M. A. (2015). Neural network prediction of ICU length of stay following cardiac surgery based on pre-incision variables. PLoS One, 10(12), e0145395.

[34] Ahmad, R. (2021). The role of digital technology and artificial intelligence in diagnosing medical images: a systematic review. Open Journal of Radiology, 11(01), 19.

[35] Aduszkiewicz, A., Ali, Y., Andronov, E., Antićić, T., Antoniou, N., Baatar, B., ... & Wojtaszek-Szwarc, A. (2017). Two-particle correlations in azimuthal angle and pseudorapidity in inelastic p+ p interactions at the CERN Super Proton Synchrotron. The European Physical Journal C, 77, 1-15.

[36] Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

[37] Trzciński, T., Graczykowski, Ł., Glinka, M., & ALICE Collaboration. (2020). Using random forest classifier for particle identification in the ALICE experiment. In *Information Technology, Systems Research, and Computational Physics 3* (pp. 3-17). Springer International Publishing.

[38] Yalcin Kuzu, S. (2022). J/ψ production with machine learning at the LHC. *The European Physical Journal Plus*, *137*(3), 392.

[39] Chen, C., Liaw, A., & Breiman, L. (2004). Using random forest to learn imbalanced data. University of California, Berkeley, 110(1-12), 24.

[40] Agusta, Z. P. (2019). Modified balanced random forest for improving imbalanced data prediction. International Journal of Advances in Intelligent Informatics, 5(1), 58-65.

[41] Khalilia, M., Chakraborty, S., & Popescu, M. (2011). Predicting disease risks from highly imbalanced data using random forest. BMC medical informatics and decision making, 11, 1-13.

[42] Amin, M., & Ali, A. (2018). Performance evaluation of supervised machine learning classifiers for predicting healthcare operational decisions. Wavy AI Research Foundation: Lahore, Pakistan, 90.

[43] Fonarow, G. C., Adams, K. F., Abraham, W. T., Yancy, C. W., Boscardin, W. J., & ADHERE Scientific Advisory Committee. (2005). Risk stratification for in-hospital mortality in acutely decompensated heart failure: classification and regression tree analysis. Jama, 293(5), 572-580.

[44] Peterson PN, Rumsfeld JS, Liang L, et al. A validated risk score for in-hospital mortality in patients with heart failure from the American heart association get with the guidelines program. Circ Cardiovasc Qual Outcomes 2010;3:25–32.

[45] Kipnis, E., Ramsingh, D., Bhargava, M., Dincer, E., Cannesson, M., Broccard, A., ... & Thibault, R. (2012). Monitoring in the intensive care. Critical care research and practice, 2012.

[46] Wang, N., Gallagher, R., Sze, D., Hales, S., & Tofler, G. (2019). Predictors of frequent readmissions in patients with heart failure. *Heart, Lung and Circulation*, *28*(2), 277-283.

[47] Web ref: Zhou, Jingmin et al. (2021), Prediction model of in-hospital mortality in intensive care unit patients with heart failure: machine learning-based, retrospective analysis of the MIMIC-III database, Dryad, Dataset, https://doi.org/10.5061/dryad.0p2ngf1zd

[48] Bennett, N. (2021). Enabling External Validation for Machine Learning Applications Using Intensive Care Data (Doctoral dissertation, ETH Zurich).

[49] Probst, P., & Boulesteix, A. L. (2017). To tune or not to tune the number of trees in random forest. The Journal of Machine Learning Research, 18(1), 6673-6690.

[50] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.

[51] Narsky, I., Porter, F.C., (2014). Statistical Analysis Techniques in Particle Physics, Almanya:Wiley–VCH.

[52] Jeni, L. A., Cohn, J. F., & De La Torre, F. (2013, September). Facing imbalanced data--recommendations for the use of performance metrics. In 2013 Humaine association conference on affective computing and intelligent interaction (pp. 245-251). IEEE.

[53] Bauder, R., & Khoshgoftaar, T. (2018, July). Medicare fraud detection using random forest with class imbalanced big data. In 2018 IEEE international conference on information reuse and integration (IRI) (pp. 80-87). IEEE.

[54] Müller A. C., Guido, S., (2016). Introduction to Machine Learning with Python, O'Reilly Media Inc., Amerika: Sebastopol Kaliforniya.

[55] Ilhan Taskin, Z., Yildirak, K., & Aladag, C. H. (2023). An enhanced random forest approach using CoClust clustering: MIMIC-III and SMS spam collection application. Journal of Big Data, 10(1), 38.

[56] Zhu, M., Xia, J., Jin, X., Yan, M., Cai, G., Yan, J., & Ning, G. (2018). Class weights random forest algorithm for processing class imbalanced medical data. IEEE Access, 6, 4641-4652.