SOCIETY FOR SCIENCE AND EDUCATION
UNITED KINGDOM

TMLAI TRANSACTIONS ON MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

VOLUME 4, ISSUE 3
ISSN 2054-7390

Comparison of Predictive Models for Transferring Stroke In-Patients to Intensive Care Unit

Nawal N. Alotaibi and Sreela Sasi
Computer and Information Science, Gannon University Erie, PA, USA
alotaibi014@gannon.edu, sasi001@gannon.edu

ABSTRACT

The cost for Intensive Care Unit (ICU) resources is extremely high and it affects healthcare budget that provides quality healthcare service for patients. Thus, the need for a predictive model for the decision to transfer stroke in-patients to the ICU is very important to utilize the resources effectively. Also, it will help to lower morbidity and mortality rates through earlier detection and intervention. In this research, initially, a Decision Tree (DT) model, an Artificial Neural Network (ANN) model, a Support Vector Machine (SVM) model, and a Logistic Regression (LR) model are evaluated for predicting the need to transfer the stroke in-patients to the ICU or not. The study is conducted on a clinical dataset consisting of 1,415 observations using the vital signs with six variables. This original dataset was having data imbalance and hence the result was misleading. In order to overcome this situation the Synthetic Minority Oversampling Technique was applied on the dataset. A DT model, an ANN model, a SVM model, and a LR model are evaluated again on the balanced dataset for prediction. Tree-based ensemble approaches such as Generalized Boosted Model, Adaptive Boosting (AdaBoost.M1), Random Forest and Bagged AdaBoost (Adabag) are used to improve the accuracy and performance of models. These methods were trained and tested on the balanced stroke in-patients dataset. The boosting model, AdaBoost.M1, and bagging model, Random Forest, achieved better accuracy compared to the other models. Therefore, these two models could be used for helping healthcare professionals in decision-making.

Keywords: Predictive Model; Artificial Neural Network; Support Vector Machine; Decision Tree; Generalized Boosted Model; Random Forest;

1 Introduction

Medical data has become electronic, which can be easy to manipulate and process. This data and information have become the basis for healthcare development through building predictive models for early risk prevention. Developing an early predictive model from the available medical data is a significant area of research nowadays due to its effects on the decision-making process for diagnostics and to predict a treatment plan. Predictive models have become more related to the daily operations in the healthcare field. This would help to reduce the cost of healthcare and to improve the quality of service. As predictive models have become more pervasive, the need for a generalized standard for model deployment is more significant [1].

DOI: 10.14738/tmlai.43.2051
Publication Date: 18th June, 2016
URL: http://dx.doi.org/10.14738/tmlai.43.2051

The term "diagnostics" indicates determining the specific condition about a patient's status, while the term "prognostics" indicates the forecasting about patient's state in the future based on available data [2]. This research will focus on prognostics (prediction) of future short-term conditions based on data mining techniques that make such predictions possible.

In clinical practice, there are many publications in the literature that have proposed predictive models by combining the advantages of data mining techniques. Currently, Logistic Regression (LR) and Artificial Neural Networks (ANN) are the most widely used models in biomedicine. The statistics of the number of publications indexed in Medline are 28,500 for LR, 8,500 for ANN, 1,300 for K-Nearest Neighbors (K-NN), 1,100 for Decision Tree (DT), and 100 for Support Vector Machines (SVM) [3].

The common data mining techniques that are used in healthcare are LR, DT, and ANN. Many researchers agreed that ANN is the best data mining technique employed for their study [4]. Abd Rahman et al. compared various predictive models to predict the survival of cardiac surgery patients. They used the dataset that consists of 5,154 observations with 23 variables. They developed predictive models using three data mining techniques, namely: LR, DT, and ANN. They found that ANN is the best predictive model with classification accuracy, sensitivity, and specificity [4].

The nature of the dataset is a significant issue with data mining research, especially in classification analysis. The main problem related to the nature of the current dataset is that it is an imbalanced dataset. Imbalance problems in binary classification mean that the range between two classes has a large difference in the number of samples. The class with a fewer portion of samples is called a minority class, while the one with a higher portion of samples is called a majority class. This problem leads to misclassification when using many of the classification algorithms because these algorithms focus on the majority class rather than the minority class. Hence, the classification result will be biased towards the majority class. As a result, this will produce a strong (classifier) predictive model that has a high predictive accuracy for the majority class and weak (classifier) predictive model for the minority class.

The imbalanced dataset problem frequently occurs in the medical diagnosis for rare diseases, but it plays an important role while deciding the treatment. Therefore, the misclassification problem for rare diseases will result in high health risks. Sampling is the main solution for imbalanced dataset problems. There are two types of sampling techniques: undersampling and oversampling. Undersampling will randomly delete majority class sample. Oversampling will reproduce minority class sample. Both methods help to achieve a balanced class distribution [5].

The ensemble method is one of the most powerful techniques in data mining analysis. It combines multiple models into a single predictive model that is often more accurate and reliable than its components. These methods provide a significant proof that supports decision-making for most of the predictive systems such as weather forecasting, market analysis, product recommendations, and medical prognostics.

The widely recognized feature of ensemble methods is its capability to combine various classifiers to achieve a new classifier that outperforms every one of them for greater performance. This is done by integrating weak classifiers to produce a strong classifier. Hence, the basic idea is to build several classifiers from the original dataset and then aggregate their predictions when new unknown samples are added.

There are two primary types of ensemble methods: boosting [6] and bagging [7]. The boosting method selects a subset of the original dataset without replacement for training each model. It focuses on the training data that is incorrectly classified at each iteration. Originally, the training dataset has equal weight. When the training dataset is incorrectly classified then its weight will be increased. Voting at each iteration will produce the final prediction model outcome. The bagging method selects a random subset of the original dataset with replacement for training each model. It uses the voting scheme to predict the final model outcome.

This paper is organized as follows: Section 2 provides literature review about predictive models. The architecture of research is explained in Section 3. The simulation and results are presented in Section 4, and Section 5 consists of the conclusion and future work.

2 Background Research

According to the World Health Organization's (WHO) global report, the death rate of chronic diseases, such as stroke in most countries will increase in 2015, reaching 41 million people [8]. The death rates are increasing, and the side effects associated with a stroke are becoming worse. An early warning system for predicting the deterioration would be more useful. Also, due to the busy work environment in hospitals, there is a lack of communication between healthcare professionals.

Data mining techniques help to extract valuable information from healthcare data. Prediction of patients' status is a highly sensitive issue to avoid worst complications that may lead to patient deterioration, especially due to chronic diseases. Hence, healthcare professionals need a simple, general, interpretable, and trustworthy predictive model for making a faster decision for stroke in-patients.

Many researchers tried to improve the predictive models' accuracies for reducing the cost in healthcare. Predictive models assist in making an appropriate decision about a patient's treatment plan. This can be generalized for choosing the only necessary healthcare procedures, which in turn will contribute to decrease healthcare costs.

Ghavami and Kapur proposed machine-learning techniques to provide early predictions about the future health status of patients to avoid medical interventions. They compared the accuracy of four neural network models such as Probabilistic Neural Networks (PNN), Support Vector Machine (SVM) networks, Multi-Layer Perceptron (MLP) with Levenberg-Marquardt (LM), and the Generalized Feed forward with LM. They found that the SVM algorithm was the most accurate one. This is followed by the MLP model and the General Feed forward Neural Network model in accuracy. Also, they mentioned that the algorithms' performances couldn't be generalized, due to the nature of the dataset. Therefore, some algorithms give better performance on different datasets than others because it depends on the dataset type and volume [2].

Toledo et al. developed a predictive model for patients with subarachnoid hemorrhage using machine learning techniques and extracted knowledge from the medical data. They applied DT, Nearest Neighbor with generalization, and Ripple Down Rule Learner, and they found that the DT is the most accurate one [9].

Several techniques have been proposed to solve the problems associated with the imbalanced class dataset. Within ensemble methods the most helpful technique to overcome the imbalance problem is boosting. Boosting is a great ensemble-learning algorithm that improves the performance of the weak classifier. The algorithms such as RUSBoost and SMOTEBoost are examples of boosting algorithms [10]. Many solutions have been proposed to overcome imbalanced dataset problems. Nguyen et al. used an oversampling technique with SVM predictive model for the data imbalance problem [11]. Their method provides an effective improvement on the dataset.

A number of undersampling methods have also been recommended to solve the imbalanced dataset problem. The undersampling method can lead to loss of some very important samples of the dataset because it focuses on randomly eliminating some samples of the majority class. Recently, advanced undersampling techniques, such as EasyEnsemble and BalanceCascade are presented in [12]. Another approach that was used to solve imbalance class is Synthetic Minority Oversampling Technique (SMOTE). Chawla et al. combined undersampling method with SMOTE [13] that has achieved better classifier performance.

In ensemble literature, Researchers presented many methods for improved model performance using the ensemble techniques. Opitz and Macklin presented a comparison and evaluation of boosting and bagging on multiple datasets [14]. They showed a comprehensive experimental study of these methods on data analysis. Freud et al. presented an overview in about boosting that works to obtain a strong classifier [15].

The main reason for using ensemble methods is to improve the performance of model. Thus, the ensemble outcome will be the optimal model for classification or prediction. The main motivation for the combination of classifiers in the ensemble is to improve their generalization ability as well. Rokach showed a review of existing ensemble methods as a tutorial for building ensemble based systems [16].

Tree-based ensemble classification models are a subset of the commonly employed ensemble modeling. There are many tree-based ensemble methods used to solve most classification and prediction problems such as random forest model. Ensemble tree techniques are appropriate for a dataset because they provide the visual analysis of predictors' participation in the final model outcome.

This research focuses on a predictive model for transferring stroke in-patients to the ICU using data mining techniques. Predictive modeling is the process of building a classifier model for a target variable using a training dataset. It is then tested on a test dataset and is validated on a validation dataset. The DT, ANN, SVM, and LR are evaluated on the original dataset in [17]. The dataset used was having data imbalance problem and hence the result was misleading. This led to having a high value of prediction accuracy because the classification was mainly focusing on majority class. Therefore, this research is explored sampling approach to overcome data imbalance problem. Then the DT, ANN, SVM, and LR are evaluated again using the sampled dataset. In addition, this research provides a comparison of tree-based ensemble models for predicting the transfer of stroke in-patients to the ICU or not. Main boosting methods such as Generalized Boosted Model (GBM) [18] and Adaptive Boosting (AdaBoost.M1) [19] were applied to the balanced stroke in-patients dataset. Also, the main bagging methods such as Random Forest [20] and Bagged Adaboost (Adabag) [21] were applied to the same dataset to improve the model accuracy. This would help the healthcare professionals to take early actions for avoiding stroke patients' deterioration as much as possible. The architecture for predictive model for transferring stroke in-patients to the ICU (PMT-ICU) is explained in Section 3.

3 Architecture of PMT-ICU

The architecture for PMT-ICU is shown in Figure 1. Data mining approaches are explored for modeling the stroke in-patients' dataset. Pre-processing is done to understand the dataset, which includes variable identification, data type conversion, and a process to deal with missing data. The dataset was modeled using the DT, ANN, SVM, and LR models. The models are evaluated to predict the ICU transfer of stroke in-patients. Classification measurements such as accuracy, sensitivity, and specificity were used to evaluate these models. Since the dataset used has an imbalanced class problem, the sampling technique is used. Synthetic Minority Oversampling Technique (SMOTE) was applied to the dataset to produce a more balanced dataset. The DT, ANN, SVM, and LR models are evaluated again using balanced dataset to predict the ICU transfer of stroke in-patients. Classification measurements such as accuracy, sensitivity, and specificity were used to evaluate these models. The tree-based ensemble approach is explored for modeling the decision for transferring stroke in-patients' to the ICU using a sampled stroke in-patients' dataset. GBM and AdaBoost.M1 from boosting approach were applied to this dataset. Also, Random Forest and Adabag from bagging approach were applied to the same dataset in order to improve model accuracy. The classification measurements such as accuracy, kappa, and Area Under the Curve (AUC) were used to evaluate these models.

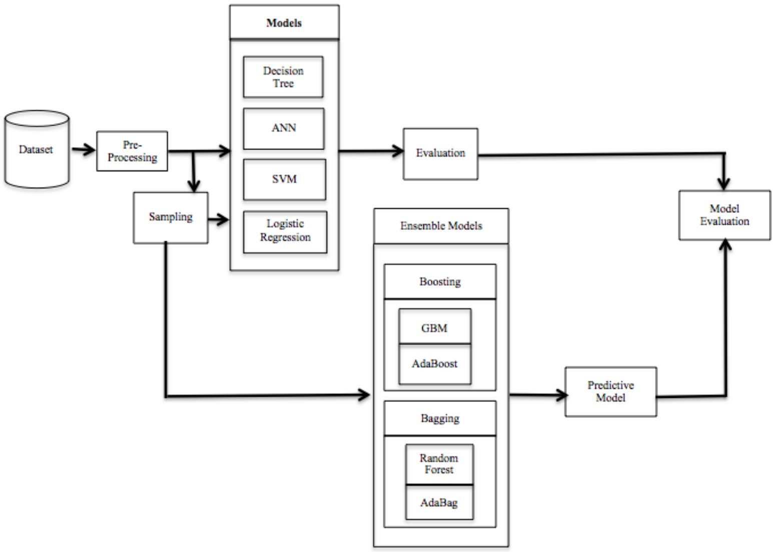


Figure 1. Architecture of Predictive Model for Transferring Stroke In-Patients to the ICU

3.1 Data Modeling

Decision tree (DT) works well with numerical and categorical variables and is easy to interpret. Chang et al. used the DT model for recognizing skin diseases [22]. Researchers in the healthcare area commonly use DT. In this research, the DT model is used to decide whether stroke in-patients need to be transferred to the ICU or not. This is done by splitting the dataset for training, testing, and validation subsets in the ratio of 70:20:10. The resulting tree obtained from the training set has six levels and fourteen leaves; so it is a complex one and is difficult to interpret. This may lead to poor performance because of the high number of levels, branches, and leaves. This is shown in Figure 2

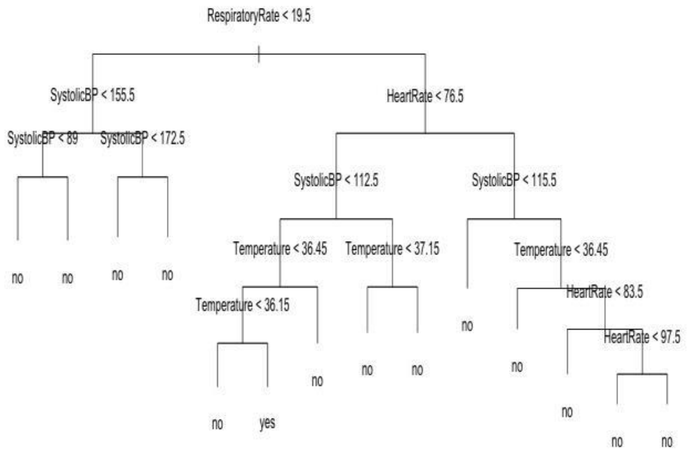


Figure 2. Decision Tree-based Classification

A smaller tree with fewer branches often leads to less complexity, easier interpretation, and minimal test and validation error rates. This solution is achieved by pruning the tree using the cross-validation function; the cv.tree in the 'tree' package in the R program is used. This function is used to choose the complexity level of the DT tree. Pruning is a technique that reduces the size of a DT by removing the sections of the tree that contribute significantly less for classification. Hence, this will reduce the complexity and improve the accuracy. Cross-validation is a model validation technique for assessing the accuracy of the result of a prediction. The cross-validation graph for pruning the DT is shown in Figure 3.

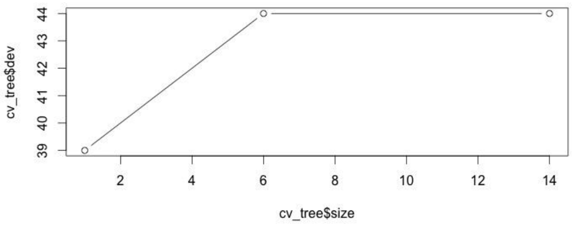


Figure 3. Cross-validation graph for pruning the Decision Tree

The pruned DT has only six leaves, as shown in Figure 4. This tree is used for testing and validation on datasets so that the model effectiveness could be checked. Then the percentage of accuracy of the model could be computed.

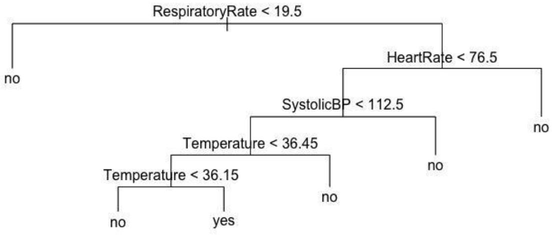


Figure 4. Pruned Decision Tree

Uttreshwar and Ghatol used a generalized regression neural network for predicting Hepatitis B, and they achieved high classification accuracy [23]. ANN is generally used for classification and pattern recognition. It is used to find and learn a hidden pattern from the dataset that could be used to build a classifier model. This model can be used for classification of a new pattern. In this research, the ANN is trained, tested and validated using the same ratio of 70:20:10 on the stroke in-patients dataset. A single hidden layer neural network based on back propagation algorithm is used for training the dataset and is shown in Figure 5. The inputs are temperature, respiratory rate, heart rate, systolic blood pressure, and oxygen saturations, and the output is ‘Yes’ or ‘No’ corresponding to whether the stroke in-patients needs to be transferred to the ICU or not. Testing and validation are done on the model to check the performance.

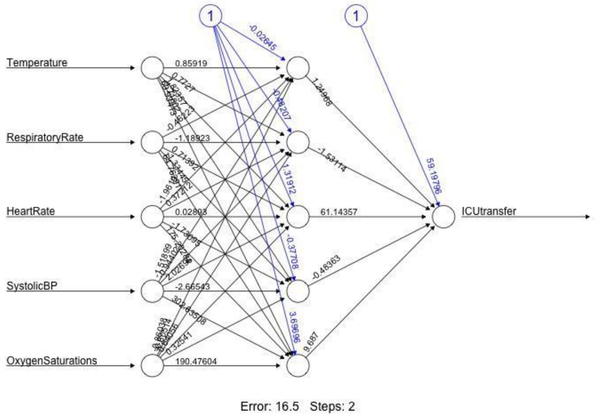


Figure 5. The Artificial Neural Network structure

Lung Chan et al. used SVM for predicting the mortality in general, and high accuracy was achieved for this model [24]. SVM is one of the most widely used techniques in the healthcare area for classification. In this research an SVM was developed for binary classification and this corresponds to the target variable that has binary value. This target variable will be the outcome of SVM model that specifies whether the stroke in-patients need to be transferred to the ICU or not. The polynomial kernel of the SVM was trained using the training dataset and tested on the testing dataset for assessing the classifier’s performance. Also, a final validation was performed on the validation dataset to check the performance of the model.

LR is a very popular technique used in medical research for predicting patients’ outcomes. Vairavan et al. proposed an algorithm based on the LR and Hidden-Markov models for predicting the mortality in the ICU using vital sign data [25]. LR works very well for categorical variables to predict the outcome. This is also compatible with this current research to predict the transfer of stroke in-patients to the ICU or not.

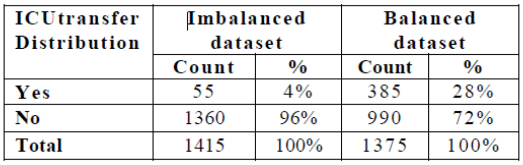
3.2 Data Sampling

Data sampling is very important in data mining analysis, especially for a large data volume. Prediction result for classification algorithms depends on the nature of dataset. If the dataset is imbalanced the classification result will be biased towards the majority class. Hence, this classifier will show the weak result on minority class. Synthetic Minority Oversampling Technique (SMOTE) is the powerful technique that used to resolve this issue.

3.2.1 Synthetic Minority Oversampling Technique (SMOTE)

The original dataset with 1,415 cases had an imbalanced dataset problem. The percentage of patients who were transferred to the ICU was only 4% (55 cases) while cases who were not transferred to the ICU was 1,360 (96%). The sampling technique was employed to overcome this problem. After the sampling was performed, the dataset had 385 (28%) cases of patients who were transferred to the ICU and 990 (72%) patients who were not, which gives a total of 1,375 patients. Table 1. shows the comparison of class distribution in imbalanced and balanced datasets.

Table 1. Comparison of Class Distribution



ICUtransfer Distribution	Imbalanced dataset		Balanced dataset	
	Count	%	Count	%
Yes	55	4%	385	28%
No	1360	96%	990	72%
Total	1415	100%	1375	100%

SMOTE generates artificial samples of the minority class. The first task of this technique is oversampling the minority class based on the nearest neighbors. Then it is used to undersample the majority class based on the number of new artificial samples that were added to the minority class during oversampling. To address this problem, SMOTE function from the DMwR package in R was used.

This function has a sequential control of the number of oversampling of the minority class and the number of undersampling of the majority classes. In this dataset, the number of new samples added to the minority class will be calculated as (600/100) = 6. Now 6 new samples will be added for each minority class of the original dataset as follows: (6*55) = 330. So the new value of minority class in the balanced dataset is 385 (330+55). The next step is the undersampling of the majority class based on the oversampling result. The number of the selected majority class for the final balanced dataset will triple the oversample of the minority class as follows: (300/100)*330 = 990 samples for the majority class in the final dataset.

3.2.2 DT, ANN, SVM, and LR Modeling on Sampled Dataset

Since the dataset is balanced, it can be used for training and testing again. Thus, the dataset is split into three portions as follows: 70% of the dataset (963 patients) for training, 20% (277 patients) for testing, and 10% (135 patients) for validation. The classification algorithms DT, ANN, SVM, and LR are run individually for training, testing, and validation using the sampled dataset.

DT works well with numerical and categorical variables and is easy to interpret. In this research, the DT model is used to decide whether stroke in-patients need to be transferred to the ICU or not. The resulting tree obtained from the training set has six levels and twelve leaves; so it is a complex one and is difficult to interpret. This may lead to poor performance because of the high number of levels, branches, and leaves. This is shown in Figure 6.

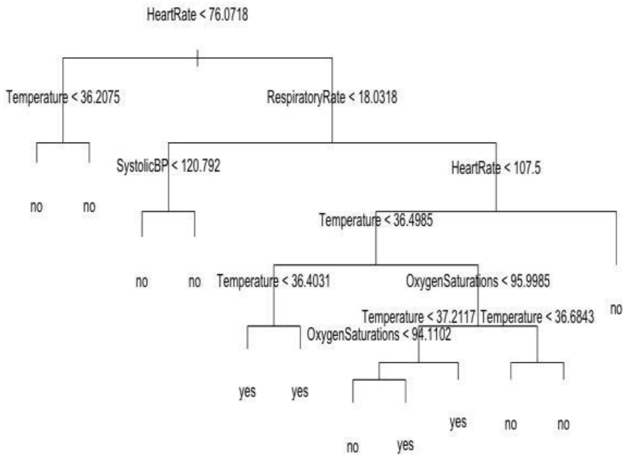


Figure 6. Decision Tree-based Classification

A smaller tree with fewer branches often leads to less complexity, easier interpretation, and minimal test and validation error rates. This solution is achieved by pruning the tree using the cross-validation. Pruning is a technique that reduces the size of a DT by removing the sections of the tree that contribute significantly less for classification. Hence, this will reduce the complexity and improve the accuracy. Cross-validation is a model validation technique for assessing the accuracy of the result of a prediction. The pruned DT has only six leaves, as shown in Figure 7. This tree is used for testing and validation on datasets so that the model effectiveness could be checked. Then the percentage of accuracy of the model could be computed**.**

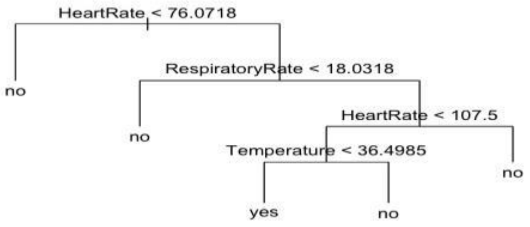


Figure 7. Pruned Decision Tree

ANN is generally used for classification and pattern recognition. It is used to find and learn a hidden pattern from the dataset that could be used to build a classifier model. This model can be used for classification of a new pattern. In this research, the ANN is trained, tested and validated using the same ratio of 70:20:10 on the stroke in-patients dataset. A single hidden layer neural network based on back propagation algorithm is used for training the dataset and is shown in Figure 8. The inputs are temperature, respiratory rate, heart rate, systolic blood pressure, and oxygen saturations, and the output is 'Yes' or 'No' corresponding to whether the stroke in-patients needs to be transferred to the ICU or not. Testing and validation are done on the model to check the performance.

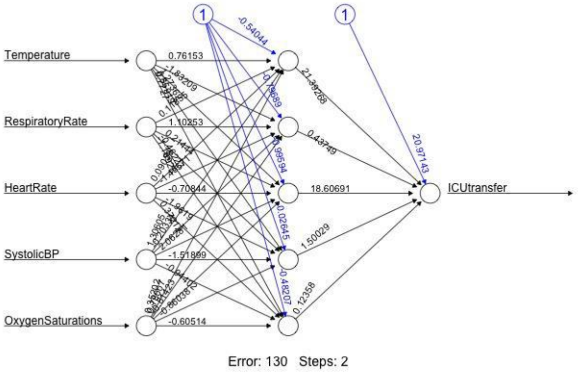


Figure 8. The Artificial Neural Network structure

SVM is one of the most widely used techniques in the healthcare area for classification. In this research an SVM was developed for binary classification and this corresponds to the target variable that has binary value. This target variable will be the outcome of SVM model that specifies whether the stroke in-patients need to be transferred to the ICU or not. The polynomial kernel of the SVM was trained using the training dataset and tested on the testing dataset for assessing the classifier's performance. Also, a final validation was performed on the validation dataset to check the performance of the model.

Logistic Regression is a very popular technique used in medical research for predicting patients' outcomes. LR works very well for categorical variables to predict the outcome. This is also compatible with this current research to predict the transfer of stroke in-patients to the ICU or not.

3.3 Tree-based Ensemble Approach

The most popular methods such as boosting and bagging methods in tree-based ensemble approach are applied on the sampled dataset. This will provide a better accuracy for the model.

3.3.1 Boosting

Boosting consists of a series of iterations for learning on the training dataset and designing samples based on the probability distribution. This method focuses on learning from incorrectly classified samples. Gradient Boosting and AdaBoost are the most common examples of boosting methods. Boosting is generally used to learn multiple models from training dataset. A weighted average is calculated using these narrowly tuned multiple models to get a good predicted result.

3.3.1.1 Generalized Boosted Model

Generalized Boosted Model (GBM) does binary classification and regression modeling. Only, binary classification is used in this research. It uses boosted trees during the iterations. GBM function uses 3–fold cross-validation of the data. This means that the model trains three times on three different subsets of the training dataset to find the best tuning model. The optimal model was produced for 150 iterations with a maximum tree depth of 3 as shown in Figure 9.

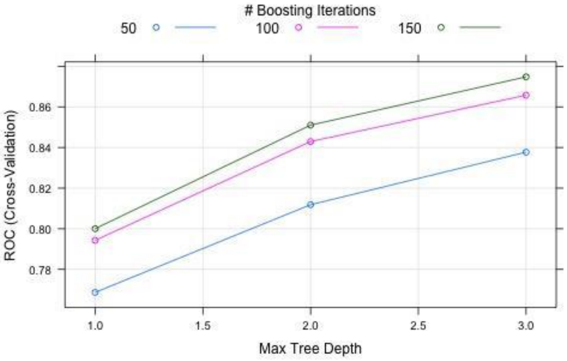
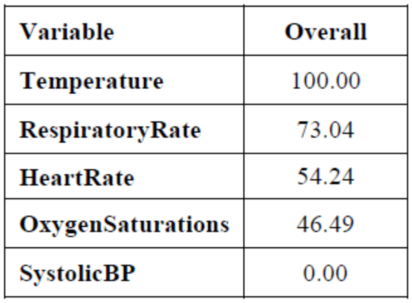


Figure 9. AUC variations by Maximum Tree Depth and number of boosting iterations for GBM

After more than hundred iterations the important variables in this model are produced and are shown in Table 2 and is also presented as a graph in Figure 10. The Temperature, RespiratoryRate, HeartRate, OxygenSaturations, SystolicBP predictors have importance in descending order for the GBM model. SystolicBP is found to be of no importance.

Table 2. GBM Variable Importance and their Overall Percentage



Variable	Overall
Temperature	100.00
RespiratoryRate	73.04
HeartRate	54.24
OxygenSaturations	46.49
SystolicBP	0.00

Figure 10. GBM variable importance

3.3.1.2 Adaptive Boosting

The most common method of boosting is the adaptive boosting model (AdaBoost.M1). It adapts to any given data without specifying any particular parameters. This model can be used for both classification and regression problems. AdaBoost.M1 trains multiple subsets for classification of dataset and assigns equal weights to each subset. It provides a number of weak classifiers on different weighted subsets of the training data. After multiple iteration of training the models it will find whether the dataset is incorrectly classified or not. Then a higher weight will be added to the model that has an incorrect classification. Figure 11 shows the relation between the 'number of trees' and 'the maximum depth' for all the three options of AdaBoost.M1 implementations (Breiman, Freund, and Zhu).

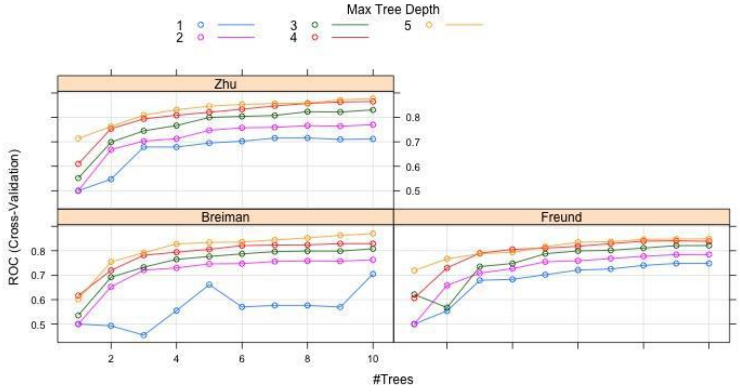


Figure 11. AUC variations by Maximum Tree Depth and number of boosting iterations for AdaBoost.M1

The HeartRate, RespiratoryRate, Temperature, OxygenSaturations, SystolicBP predictors have importance in descending order for the AdaBoost.M1 model as shown in Table 3 and is also presented as a graph in Figure 12. SystolicBP is found to be of no importance.

Table 3. Adaboost.M1 Variable Importance and their Overall Percentage

Variable	Overall
HeartRate	100.00
RespiratoryRate	69.42
Temperature	46.37
OxygenSaturations	40.88
SystolicBP	0.00

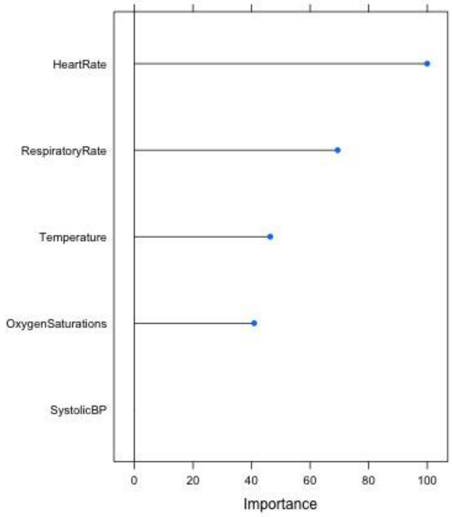


Figure 12. AdaBoost.M1 variable importance

3.3.2 Bagging

Bagging (Bootstrap Aggregation) is designing samples from the training dataset with replacement, and then getting the multiple models to learn from different random samples of the training dataset. Separate classifier is used to train these different samples. When a new data comes all these classifiers will be used. Then a voting scheme is used to predict the outcome based on these classifiers. The highest vote obtained using these classifiers is the final outcome used for classification. In the regression case, the average of estimated output from all these classifiers is taken. The main example of bagging is Random Forest. Bagging is used to reduce the variance of the prediction through creating extra samples for training dataset with replacement. It focuses on decreasing the variance instead of increasing the size of training dataset.

3.3.2.1 Random Forest

Random Forest is one of the most common bagging models. This is an ensemble method that fits a number of single decision tree classifiers on multiple subsets of the training dataset. It applies the averaging method to improve the prediction accuracy. It builds a large number of Classification And Regression Trees (CART) [10]. When a new sample is added, each tree votes for prediction. All trees in the forest participate in voting for the final result. The variable that gets the majority of the votes will be the final classification result. Random Forest uses k-cross validation model to select the proper model parameters. These tuning parameters are the number of randomly selected predictors in each split of the tree. The optimal model is produced with AUC =0.93 using 2 random predictors as given in Figure 13.

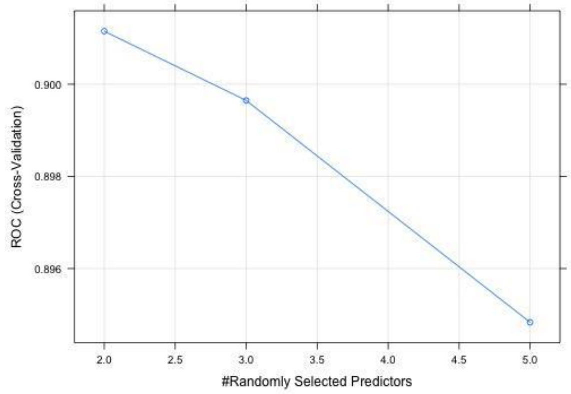


Figure 13. Area-Under-Curve scores versus number of selected predictors randomly for Random Forest

The HeartRate, Temperature, RespiratoryRate, SystolicBP, OxygenSaturations predictors have importance in descending order for the Random Forest model as shown in Table 4 and is also presented in Figure 14 as a graph. OxygenSaturations is found to be of no importance.

Table 4. Random Forest Variable Importance and their Overall Percentage

Variable	Overall
HeartRate	100.00
Temperature	87.73
RespiratoryRate	76.56
SystolicBP	44.64
OxygenSaturations	0.00

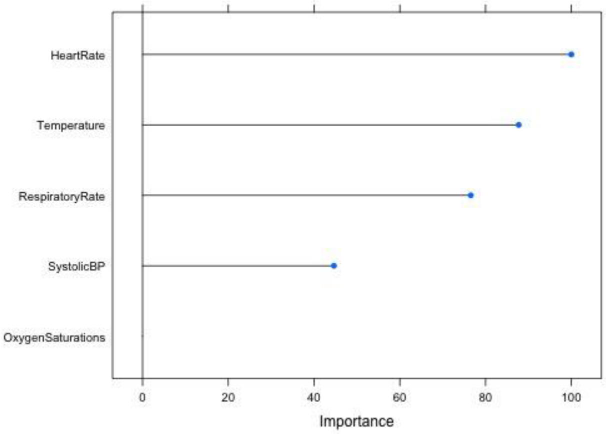


Figure 14. Random Forest variable importance

3.3.2.2 Bagged AdaBoost

Bagged AdaBoost (Adabag) is designed for using classification trees as individual classifiers. After these classifiers have been trained for number of iterations, they can be used for predicting a new sample of data. Adabag uses k-cross validation model to select the proper model parameters. Adabag is implemented with the features of the boosting and bagging classification methods that use a tree as base classifier. The optimal model was produced using the number of trees=10 with maximum tree depth = 5 as shown in Figure 15.

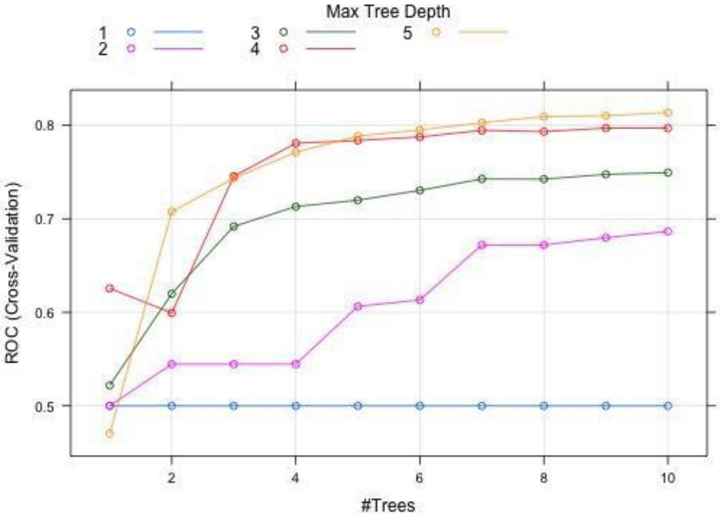


Figure 15. AUC variations by Maximum Tree Depth and number of boosting iterations for Adabag

The Temperature, HeartRate, RespiratoryRate, SystolicBP, OxygenSaturations predictors have importance in descending order for the Adabag model as shown in Table 5, and is also presented in Figure 16 as a graph. OxygenSaturations is found to be of no importance.

Table 5. Adabag Variable Importance and their Overall Percentage

Variable	Overall
Temperature	100.00
HeartRate	63.27
RespiratoryRate	38.16
SystolicBP	21.28
OxygenSaturations	0.00

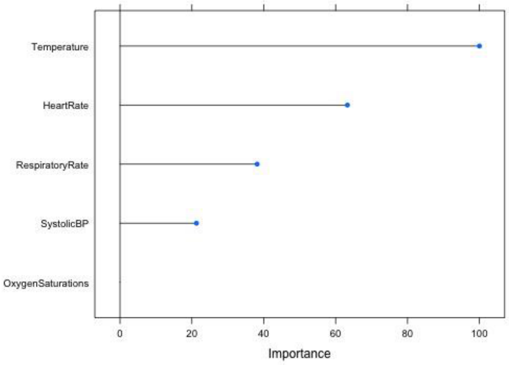
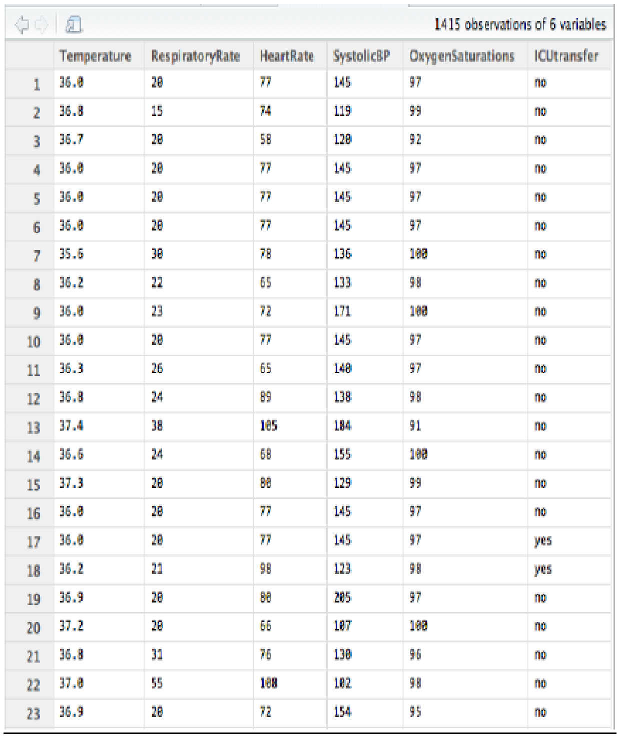


Figure 16. Adabag variable importance

4 Simulation and Results

The dataset was obtained from King Fahad Medical Center, National Neuroscience Institute, Riyadh, Saudi Arabia. A snapshot of the data is given in Table 6.

Table 6. Dataset Snapshot

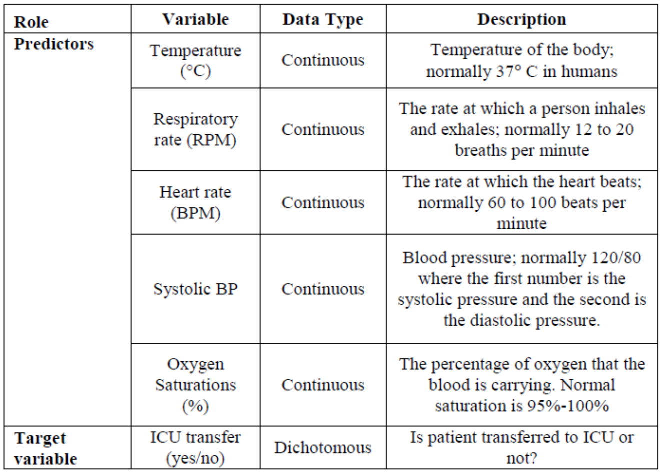


1415 observations of 6 variables

	Temperature	RespiratoryRate	HeartRate	SystolicBP	OxygenSaturations	ICUtransfer
1	36.0	20	77	145	97	no
2	36.8	15	74	119	99	no
3	36.7	20	58	120	92	no
4	36.0	20	77	145	97	no
5	36.0	20	77	145	97	no
6	36.0	20	77	145	97	no
7	35.6	30	78	136	100	no
8	36.2	22	65	133	98	no
9	36.0	23	72	171	100	no
10	36.0	20	77	145	97	no
11	36.3	26	65	140	97	no
12	36.8	24	89	138	98	no
13	37.4	38	105	184	91	no
14	36.6	24	68	155	100	no
15	37.3	20	80	129	99	no
16	36.0	20	77	145	97	no
17	36.0	20	77	145	97	yes
18	36.2	21	98	123	98	yes
19	36.9	20	80	205	97	no
20	37.2	20	66	107	100	no
21	36.8	31	76	130	96	no
22	37.0	55	108	102	98	no
23	36.9	20	72	154	95	no

The dataset consists of 1,415 patients that were admitted to the hospital due to stroke. This dataset consists of 5 independent variables and one dependent variable. The independent variables are the vital sign data, such as temperature, respiratory rate, heart rate, systolic blood pressure (BP), and oxygen saturations. These are continuous variables and are listed with description in Table 7 .The dependent variable was a dichotomous variable that represented the clinical outcome to specify whether a patient is transferred to the ICU or not. There were 55 patients (4%) that were positively transferred to the ICU, and 1,360 (96%) who were not.

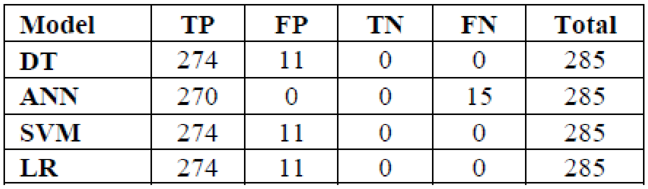
Table 7. Variables and Their Description



Role	Variable	Data Type	Description
Predictors	Temperature (°C)	Continuous	Temperature of the body; normally 37° C in humans
	Respiratory rate (RPM)	Continuous	The rate at which a person inhales and exhales; normally 12 to 20 breaths per minute
	Heart rate (BPM)	Continuous	The rate at which the heart beats; normally 60 to 100 beats per minute
	Systolic BP	Continuous	Blood pressure; normally 120/80 where the first number is the systolic pressure and the second is the diastolic pressure.
	Oxygen Saturations (%)	Continuous	The percentage of oxygen that the blood is carrying. Normal saturation is 95%-100%
Target variable	ICU transfer (yes/no)	Dichotomous	Is patient transferred to ICU or not?

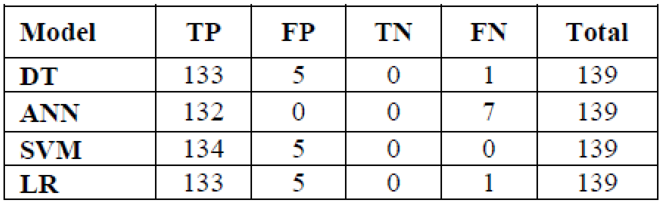
DT, ANN, SVM, and LR are individually used for training, testing, and validation using the dataset. The models were trained using 70% of the dataset (991 patients), and were tested using 20% (285 patients), and were validated using 10% (139 patients). R programming language is used for the model simulation. The simulation results of these models are compared using the classification measures, such as false positives (FP), false negatives (FN), true positives (TP) and true negatives (TN). The results for testing are given in Table 8 and for validation are shown in Table 9.

Table 8. Results Using Test Dataset



Model	TP	FP	TN	FN	Total
DT	274	11	0	0	285
ANN	270	0	0	15	285
SVM	274	11	0	0	285
LR	274	11	0	0	285

Table 9. Results Using Validation Dataset



Model	TP	FP	TN	FN	Total
DT	133	5	0	1	139
ANN	132	0	0	7	139
SVM	134	5	0	0	139
LR	133	5	0	1	139

The measurements that were used to compare the performance of these models are accuracy, sensitivity, and specificity. These measurements are expressed in equations (1), (2), and (3).

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

$$\text{Sensitivity} = \frac{TP}{TP+FN} \tag{2}$$

$$\text{Specificity} = \frac{TN}{TN+FP} \tag{3}$$

Accuracy is the fraction of true results, either true positive or true negative. Sensitivity measures the fraction of positive cases that are classified correctly as positive. Specificity is the fraction of negative cases that are classified correctly as negative. The results showed that the DT model, SVM model, and LR model have similar accuracy, which is equal to 0.96; followed by the ANN model which is 0.94. The results are shown in Table 10 for the test dataset, and Table 11 for the validation dataset.

Table 10. Classification Measurements for Test Dataset

Measurement	DT	ANN	SVM	LR
Accuracy	0.96	0.94	0.96	0.96
Sensitivity	1	0.94	1	1
Specificity	0	0	0	0

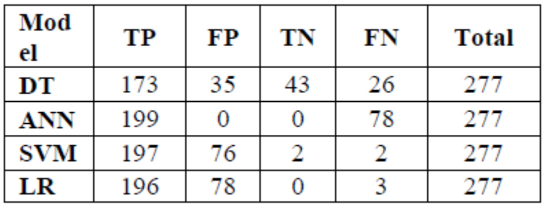
Table 11. Classification Measurements for Validation Dataset

Measurement	DT	ANN	SVM	LR
Accuracy	0.95	0.94	0.96	0.96
Sensitivity	1	0.94	1	1
Specificity	0	0	0	0

All four models showed low specificity measures that indicate the weak ability of detecting a true negative. This bias is due to the lower number of negative ICU transfer cases in this research (only 55 out of 1,415 cases). The above tables illustrate that the prediction results are not as accurate as one would like. This is due to the nature of the dataset which is original from the hospital. It is true that not many patients will be admitted to the ICU. There were very few negative cases in the entire population to train a more accurate predictive model. Also, several of the input variables were highly correlated such that the prediction contributions of some variables were less significant for making a more accurate prediction. The ideal training dataset would include an approximately equal number of positive and negative cases.

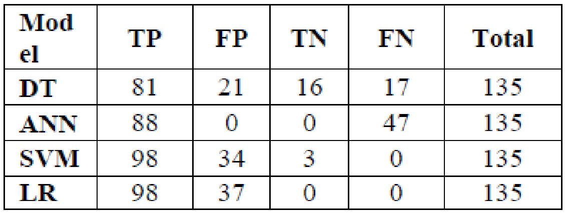
Classification accuracy is usually the primary measure used to evaluate models on classification problems. Therefore, the reason to get more than 90% of accuracy on an imbalanced dataset is because one class has more than 90% of the samples compared to the other class. In order to deal with the imbalanced dataset, sampling is done. DT, ANN, SVM, and LR are modeled using the same dataset (balanced) and they are individually used for training, testing, and validation using the sampled dataset. The models were trained using 70% of the dataset (963 patients), and were tested using 20% (277 patients), and were validated using 10% (135 patients). R programming language is used for the model simulation. The simulation results of these models are compared using the classification measures, such as False Positives (FP), False Negatives (FN), True Positives (TP) and True Negatives (TN). The results for testing are given in Table 12 and for validation are shown in Table 13.

Table 12. Results Using Balanced Test Dataset



Model	TP	FP	TN	FN	Total
DT	173	35	43	26	277
ANN	199	0	0	78	277
SVM	197	76	2	2	277
LR	196	78	0	3	277

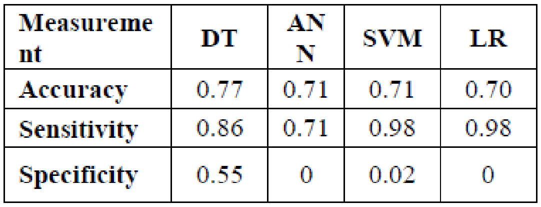
Table 13. Results Using Balanced Validation Dataset



Model	TP	FP	TN	FN	Total
DT	81	21	16	17	135
ANN	88	0	0	47	135
SVM	98	34	3	0	135
LR	98	37	0	0	135

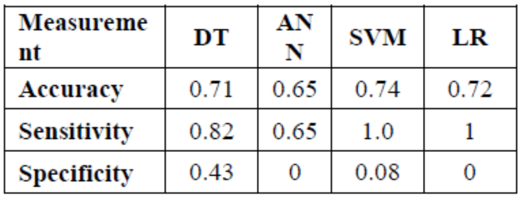
Accuracy is the fraction of true results, either true positive or true negative. Sensitivity measures the fraction of positive cases that are classified correctly as positive. Specificity is the fraction of negative cases that are classified correctly as negative. The results showed that the DT model has the best accuracy, which is equal to 0.77, followed by the ANN, SVM, and LR models that have similar accuracy around 0.71. The results are shown in Table 14 for the test dataset, and Table 15 for the validation dataset.

Table 14. Classification Measurements for Balanced Test Dataset



Measurement	DT	ANN	SVM	LR
Accuracy	0.77	0.71	0.71	0.70
Sensitivity	0.86	0.71	0.98	0.98
Specificity	0.55	0	0.02	0

Table 15. Classification Measurements for Balanced Validation Dataset



Measurement	DT	ANN	SVM	LR
Accuracy	0.71	0.65	0.74	0.72
Sensitivity	0.82	0.65	1.0	1
Specificity	0.43	0	0.08	0

Tree-based ensemble models predict the decision for transferring stroke in-patients to the ICU. Boosting and bagging are the two ensemble tree-based models applied in this research. GBM, AdaBoost.M1, Random Forest, and Adabag are individually used for training and testing using the balanced stroke in-patients dataset to improve the model performance. The total balanced dataset (1,375 cases) was split into training dataset (1,032) and testing dataset (343) in a ratio of 75% to 25% respectively.

GBM, AdaBoost.M1, Random Forest, and Adabag were trained using 75% of the balanced dataset and were tested using 25%. Accuracy, Kappa, and AUC are the measurements used to evaluate these ensemble models. Accuracy is the score of the prediction percentage. Kappa is a measurement of the model performance that gives agreement between model predictions and reality. The Kappa score ranges between -1 and +1. Scores above 0.8 are generally considered as a good agreement; less than zero means no agreement at all. AUC is the measurement for model performance that is often used for binary classification problems. An AUC score of 1 leads to an ideal model, while a score around 0.5 means random estimation. The comparison of the prediction accuracy and the performance of these four tree-based ensembles models are given in Table 16.

Table 16. Comparison Of GBM, Adaboost.M1, Randomforest, And Adabag Models

Models	Boosting Models		Bagging Models	
	GBM	AdaBoost.M1	Random Forest	Adabag
Accuracy	0.8542274	0.8950437	0.8950437	0.8192420
Kappa	0.6189566	0.7412731	0.7396255	0.5306555
AUC	0.9124	0.9373	0.9365	0.8421

The boosting model (AdaBoost.M1) and the bagging model (Random Forest) have high accuracy with errors less than 11% in comparison with other remaining models. Kappa resulted is 0.74 and 0.73 for AdaBoost.M1 and Random forest respectively. This indicates that these two are better predictive models. Also, AdaBoost.M1 and Random forest showed to have a better AUC score than other models that are close to the score of perfect model.

The Receiver Operating Characteristic (ROC) curve in binary classification can be used to evaluate the model performance using sensitivity and specificity. The ROC curve plots the sensitivity (true positive rate) by one minus specificity (false positive rate). The area under the ROC curve is a common measure of model performance.

AdaBoost.M1 and Random Forest have outperformed GBM, and Adabag providing a better sensitivity as presented in Figure 17. The AUC score for AdaBoost.M1, Random Forest, GBM, and Adabag are 0.9373, 0.9365, 0.9124, and 0.8421 respectively.

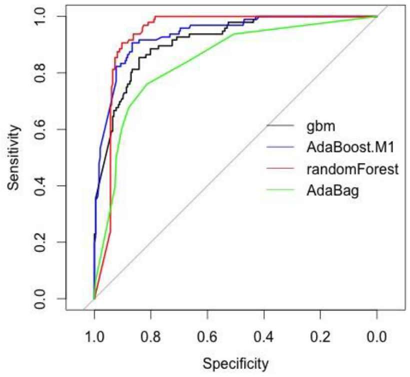


Figure 17. ROC Curves for GBM AdaBoost.M1, Random forest, and Adabag Models

5 Conclusion and Future Work

This research compared the Decision Tree, Artificial Neural Network, Support Vector Machine, and Logistic Regression models for the prediction of ICU transfer of Stroke in-patients using a clinical dataset. The comparison showed that the all four models have high and similar accuracy, due to the imbalanced class problem in the original dataset. Sampling techniques was used to overcome this problem by applying Synthetic Minority Oversampling Technique (SMOTE) on the original dataset. An evaluation of Decision Tree (DT) model, Artificial Neural Network (ANN) model, Support Vector Machine (SVM) model, and Logistic Regression (LR) model was done again on sampled dataset, and results are presented. A comparison of four tree-based ensemble models is done in this research. These are applied on a balanced stroke in-patients dataset for predicting the ICU transfer. Boosting and bagging are the two ensemble models used. Generalized Boosted Model (GBM), Adaptive Boosting (AdaBoost.M1), Random Forest, and Bagged Adaptive Boosting (Adabag) were explored. The result showed that a fair competition of boosting (AdaBoost.M1) and bagging (Random Forest) models when trained and tested on the same dataset.

This would give healthcare professionals the opportunity to perform procedures early to avoid stroke in-patients’ deterioration as much as possible. Therefore, building a clinical decision support system (CDSS) based on these results would help to make an accurate decision to whether there is a need for transferring the stroke in-patients to the ICU or not.

Future work will focus on to improve the prediction results by working with a larger data volume and a variety of datasets. Another ensemble approach could be tested to get a better predictive model such as stacking.

ACKNOWLEDGMENT

The authors would like to express their deepest gratitude to the National Neuroscience Institute at King Fahad Medical City in Riyadh, Saudi Arabia for providing the dataset for this research.

REFERENCES

[1] A. Guazzelli, 'Predictive analytics in healthcare', Ibm.com, 2015. [Online]. Available: https://www.ibm.com/developerworks/library/ba-ind-PMML3/. [Accessed: 21- Sep- 2015].

[2] P. Ghavami and K. Kapur, 'Prognostics and Prediction of Patient Health Status Using a Multi-Model Artificial Intelligence Framework', Public Health Frontier, vol. 2, no. 2, pp. 46-60, 2013.

[3] S. Dreiseitl and L. Ohno-Machado, 'Logistic regression and artificial neural network classification models: a methodology review', Journal of Biomedical Informatics, vol. 35, no. 5-6, pp. 352-359, 2002.

[4] H. Abd Rahman, Y. Wah, Z. Khairudin and N. Abdullah, 'Comparison Of Predictive Models To Predict Survival Of Cardiac Surgery Patients', Statistics in Science, Business, and Engineering (ICSSBE), 2012 International Conference on, pp. 1-5, 2012.

[5] R. Longadge, S. Dongre and L. Malik, 'Class Imbalance Problem in Data Mining: Review',International Journal of Computer Science and Network (IJCSN), vol. 2, no. 1, 2013.

[6] Y. Freund and R. Schapire, “A short introduction to boosting,” J. Japan. Soc. for Artif. Intel., vol. 14(5), pp. 771–780, 1999.

[7] L. Breiman, "Bagging predictors", Mach Learn, vol. 24, no. 2, pp. 123-140, 1996.

[8] N. Kasabov, V. Feigin, Z. Hou, Y. Chen, L. Liang, R. Krishnamurthi, M. Othman and P. Parmar, 'Evolving spiking neural networks for personalised modelling, classification and prediction of spatio-temporal patterns with a case study on stroke', Neurocomputing, vol. 134, pp. 269-279, 2014.

[9] P. de Toledo, P. Rios, A. Ledezma, A. Sanchis, J. Alen and A. Lagares, 'Predicting the Outcome of Patients With Subarachnoid Hemorrhage Using Machine Learning Techniques', IEEE Transactions on Information Technology in Biomedicine, vol. 13, no. 5, pp. 794-801, 2009.

[10] R. Longadge, S. Dongre and L. Malik, 'Class Imbalance Problem in Data Mining: Review',International Journal of Computer Science and Network (IJCSN), vol. 2, no. 1, 2013.

[11] H. Nguyen, E. Cooper and K. Kamei, 'Borderline over-sampling for imbalanced data classification', International Journal of Knowledge Engineering and Soft Data Paradigms, vol. 3, no. 1, p. 4, 2011.

[12] Xu-Ying Liu, Jianxin Wu and Zhi-Hua Zhou, 'Exploratory Undersampling for Class-Imbalance Learning', IEEE Trans. Syst., Man, Cybern. B, vol. 39, no. 2, pp. 539-550, 2009.

[13] N. Chawla, K. Bowyer, L. Hall and W. Kegelmeyer, 'SMOTE: Synthetic Minority Over-sampling Technique', Journal of Artificial Intelligence Research, vol. 16, no. 321357, 2002.

[14] D. Opitz and R. Maclin, 'Popular Ensemble Methods: An Empirical Study', Journal of Artificial Intelligence Research 11, vol. 169-198, 1999.

[15] Y. Freund and R. Schapire, “A short introduction to boosting,” J. Japan. Soc. for Artif. Intel., vol. 14(5), pp. 771–780, 1999.

[16] L. Rokach, 'Ensemble-based classifiers', Artificial Intelligence Review, vol. 33, no. 1-2, pp. 1-39, 2009.

[17] N. Alotaibi and S. Sasi, 'Predictive Model for Transferring Stroke In-Patients to Intensive Care Unit', Symposium on Machine Learning Algorithms and Applications (MLAA'15) in IEEE International Conference on Computing and Network Communications (CoCoNet - 2015), pp. 136-141, 2015

[18] J. Friedman, "Stochastic gradient boosting", Computational Statistics & Data Analysis, vol. 38, no. 4, pp. 367-378, 2002.

[19] J. Friedman, T. Hastie, and R. Tibshirani, “Additive logistic regression:a statistical view of boosting,” Ann. Statist., vol. 28, no. 2, pp. 337–407, 2000.

[20] A. Liaw and M. Wiener, "Classification and Regression by randomForest", vol. 23, 2002.

[21] E. Alfaro, M. Gámez and N. García, "adabag : An R Package for Classification with Boosting and Bagging", Journal of Statistical Software, vol. 54, no. 2, 2013.

[22] C. Chang and C. Chen, 'Applying decision tree and neural network to increase quality of dermatologic diagnosis', Expert Systems with Applications, vol. 36, no. 2, pp. 4035-4041, 2009.

[23] G. Uttreshwar and A. Ghatol, 'Hepatitis B Diagnosis Using Logical Inference and Self-Organizing Map', J. of Computer Science, vol. 4, no. 12, pp. 1042-1050, 2008.

[24] C. Lung Chan, C. Li Chen and H. Wei Ting, 'An Excellent Mortality Prediction Model Based on Support Vector Machine (SVM)-a Pilot Study', International Symposium on Computer, Communication, Control and Automation, 2010.

[25] S. Vairavan, L. Eshelman, S. Haider, A. Flower and A. Seiver, 'Prediction of Mortality in an Intensive Care Unit using Logistic Regression and a Hidden Markov Model', Computing in Cardiology, vol., no. 39-393-396, 2012