KJMS 2022 VOL.4 (2): 43-55

KDU Journal of Multidisciplinary Studies (KJMS)

Volume 4. Issue 2 Nov 2022

DOI: http://doi.org/10.4038/kjms.v4i2.50

HEART DISEASE RISK IDENTIFICATION USING MACHINE LEARNING TECHNIQUES FOR A HIGHLY IMBALANCED DATASET: A COMPARATIVE STUDY

Fernando C.D.[1] Weerasinghe P.T.[2] and Walgampaya C.K.[3]

Computing Centre, Faculty of Engineering, University of Peradeniya, Peradeniya, Sri Lanka[1]

Department of Statistics and Computer Science, Faculty of Science, University of Peradeniya, Peradeniya, Sri Lanka[2]

Department of Engineering Mathematics, Faculty of Engineering, University of Peradeniya, Peradeniya, Sri Lanka[3]

ABSTRACT

Heart disease has become one of the most prevailing universal diseases in the world today. It is estimated that 32% of all deaths worldwide are caused due to heart diseases. One of the major causes for this is that its extremely difficult even for medical practitioners to predict heart diseases as heart attacks as it is a complex task which requires a great amount of knowledge and experience. The number of deaths caused by heart diseases has hugely increased in the recent past. Machine learning has become one of the most popular areas in computer science where many complex problems have been addressed successfully specially in the field of medicine. In this study we trained multiple supervised classifiers namely'; Naïve Bayes, LightGBM, Decision Trees, Random Forest, XGBoost, K Nearest Neighbours and ADABoost and we compared the accuracies and identified what models perform better for heart disease prediction. We used the Behavioral Risk Factor Surveillance System (BRFSS) 2015 Heart Disease Health Indicators Dataset which was highly imbalanced and in order to address the class imbalance problem we used methods such as Synthetic Minority Over Sampling Technique (Smote) Sampling, Adaptive Synthetic Sampling, Random Over Sampling, Random Under Sampling, TomekLink, SmoteTomek, Smoteen and Cluster Centroid. According to the results obtained, we can conclude that the hybrid models such as Smoteen and SmoteTomek performed better than the other sampling methods.

KEYWORDS: *Heart Disease, Machine Learning, Class Imbalance, Sampling methods*

Corresponding Author: Fernando C.D. Email: *fchanna2853@gmail.com*

1. INTRODUCTION

Cardiovascular diseases (CVD's) have been the leading cause of death globally with an estimate of 17.9 million dying every year as mentioned by the World Health Organization (WHO). CVDs are a group of disorders of the heart and blood vessels and include coronary heart disease, cerebrovascular disease, rheumatic heart disease and other conditions. One of the major reasons for such an increase in the number of deaths can be caused due to unhealthy diet, physical inactivity, tobacco use and harmful use of alcohol. The behavioral risk factors might show up in people due to raised blood pressure, raised blood glucose, raised blood lipids, and overweight and obesity. The risk factors can be measured in primary care facilities and indicate an increased risk of heart attack, stroke, heart failure and other complications. Prevention of tobacco use, reduction of salt in diet, eating more fruits and vegetables, regular physical exercises and avoiding harmful use of alcohol have shown to reduce the risk of cardiovascular disease. Identifying people at highest risk of CVD's and making sure that early symptoms are detected and treated can prevent premature deaths. But, diagnosis is a major problem for practitioners as the nature of the symptom is similar to other conditions and are often confused with signs of aging.

The growth of data in the field of medicine has given new opportunities for physicians to improve patient diagnosis.

There is a great deal of interest in employing computer technologies to enhance decision support and offer solutions to challenging issues in the field of medicine, especially with the development of computer science and the quick rise of Artificial Intelligence, Machine Learning, and Computer Vision. Machine learning has become one of the most popular methods used to diagnose, detect and forecast many cardiovascular disorders. This has given the opportunity to identify new ways to predict heart diseases by detecting and treating symptoms at an early stage. In this study, we train multiple classifiers both supervised and unsupervised and compare the accuracies to identify which machine learning algorithms perform better in heart disease detection.

The main contributions of our study are given below:

- Address the class imbalance problem of real world medical datasets by applying many statistical methods on sampling techniques in order to obtain higher accuracies for our models.
- Identify which classification algorithms are suitable for corresponding statistical sampling methods.
- Identifying the statistical sampling methods used to address the class imbalance problem that are able to obtain the highest accuracies.

2. LITERATURE REVIEW

Boshra Bahrami (2015) evaluated a standard dataset acquired from a hospital in Iran containing 209 records and 8 features with different machine learning classifiers in heart disease diagnosis. Since this dataset doesn't contain missing values, they have straight away selected features using both Gain Ratio Attribute Evaluated and Ranker Search methods. Then using the WEKA data mining tool, the data set was applied to four machine learning algorithms by using 10-fold cross-validation. (J48 Decision Tree classifier, K Nearest Neighbor classifier, Support Vector Machine classifier, and Naive Bayes Classifier) Finally, they measured the accuracy, precision, sensitivity, specificity, F-measure, and the area under Receiver Characteristic Operator (ROC) curve. The J48 Decision Tree classifier showed the best results by achieving higher values for the above-mentioned parameters.

Asha Rajkumar (2022) employed the "Tangara" data mining tool which comes in handy with a graphical user interface to compare the performance of selected supervised machine learning algorithms. The training dataset consists of 3000 samples and 14 features for each instance. Naive Bayes Classifier, K Nearest Neighbor classifier, and Decision List algorithms were applied to the above data set using the Tangara data mining tool. As for the analysis results, the Naive

Bayes algorithm showed the best performance in terms of the computation time, the accuracy, the left ventricle hypothesis, the normal and the stress abnormal statistical parameters.

Choosing the best features out of the dataset is one of the important facts that decide the accuracy of the data analysis results Jian Ping L (2021). In this study he proposed a fast conditional mutual feature selection algorithm (FCMIM) to select the best features out of the raw dataset. In this method, by applying Conditional Mutual Information (CMI), each feature is given a score with respect to the output class and other features and selects the best features out of them. The newly developed feature selection method and existing feature selection methods (Relief Algorithm, Minimal Redundancy Maximal Relevance Algorithm, Least Absolute Shrinkage Selection Operator Algorithm, Local Learning Based Features Selection Algorithm) were applied to the "Cleveland Heart Disease "dataset which consists of 303 instances with 75 features. The preprocessed outputs of each algorithm are then fed to selected machine learning classifiers and calculated by several statistical parameters to compare the performance of each feature selection method. The Proposed feature selection method (FCMIM) with Support Vector Machine Classifier showed the highest accuracy compared to the other mentioned algorithms.

Selecting the best features and balancing the output classes can greatly improve the data analysis results of a dataset. Abid Ishaq (2020) used a heart failure dataset to improve the heart disease patient survivor prediction. They employed Decision Tree (DT), Adaptive boosting classifier (AdaBoost), Logistic Regression (LR), Stochastic Gradient classifier (SGD), Random Forest (RF), Gradient Boosting classifier (GBM), Extra Tree Classifier (ETC), Gaussian Naive Bayes classifier (G-NB), and Support Vector Machine (SVM) machine learning algorithms and compared output results. In the proposed approach, the Random Forest Algorithm is used to select features and the Synthetic Minority Oversampling Technique (SMOTE) is used to balance the output class and then analyzed the outputs of the selected machine learning algorithms. In comparison to the original data set, the findings demonstrated that the ETC classifier had the maximum accuracy when the suggested data pretreatment technique was used.

Norma Latif Fitriyani (2020) proposed an Effective Heart Disease Prediction Model (HDPM) as a Clinical Decision Support System (CDSS) to identify heart diseases in the earlier stages. In this system, they used a Density-Based Spatial Clustering of Applications with Noise (DBSCAN) to remove outliers from the dataset. Then applied Synthetic Minority Over-sampling Technique-Edited Nearest Neighbor (SMOTE-ENN) algorithm to solve the class imbalance problem. As for the data analysis algorithm, they employed Extreme Gradient Boosting (XGBoost) to train the dataset. Applying the proposed system with the "Statlog" and the "Cleveland" dataset they achieved 95.90% and 98.40% accuracies respectively. To help medical professionals better diagnose patients, this proposed system has been implemented as a Heart Disease Clinical Decision Support System (HDCDSS).

3. METHODOLOGY

Behavioral Risk Factor Surveillance System (BRFSS) is one of the major surveys conducted in the United States of America (USA). BRFSS interviewed around 400,000 adults each year and remotely over the telephone. The questions are based on health-related risk behaviors, chronic health conditions, and the use of preventive services. This survey is conducted in all states of U.S.A., the District of Columbia, and three other U.S. territories (Snead 2020).

The data set used in this study is the data from the 2015 BRFSS report. The original data contains 441,455 responses regarding risk factors for heart disease with 330 features. The cleaned data set contains 253,680 instances with 21 quantitative features. This data set doesn't contain any missing values. The output is a binary class containing 229,787 responses from people who didn't have heart disease and the rest of 23,893 responses from those who suffered from heart disease.

The features of this dataset are given in table 2below. The following are the final outputs of the dataset.

0 = No high blood pressure

1 = High blood pressure

Class Imbalance

A data set is considered "imbalanced" when the output classes show a skew distribution. The output classes with a higher number of samples are called "majority classes" while the classes which have a fewer samples are called "minority classes". But this class imbalance does not affect the results of the classification algorithms (M Galar 2012).

In this dataset, the number of people who haven't suffered from heart disease is very high compared to the people who have. The ratio between those two is 9.62%.

Table 1: Dataset Output Classes

Category	Number of Samples
Had not suffered from heart disease (0)	229,787
Suffered from heart disease (1)	23,893

Since the machine learning classification algorithms expect a balance between minority and majority classes, this class imbalance has to be addressed before fetching the dataset into any machine learning algorithm. If the class imbalance is not addressed properly, the classification algorithms will show higher accuracy for the majority classes and poor accuracy for the minority classes.

The results of the algorithm will be biased towards the majority class, while the minority class will be almost neglected (Sing A. 2015). There are mainly three methods to solve class imbalance.

- Over Sampling.
- Under Sampling.
- Hybrid Sampling (Combination of Over Sampling and Under Sampling).

Over Sampling.

Oversampling methods try to increase the number of samples in the minority class by adding new synthetic samples into the minority class. This will improve the ratio between minority and majority classes thus balancing both classes equally. To achieve this, there are a lot of methods that are being used in many studies. Some of the techniques just replicate existing samples and balance classes and some techniques generate synthetic samples by creating new samples with different strategies (Rančić S. 2021). These oversampling methods can help to improve the performance of the machine learning models (Kovács G. 2019). In this study, we have used the following over sampling techniques on the dataset.

ROS: Random Over Sampling.

This is a simple process to increase the minority class size by duplicating randomly selected data samples from the minority class depending on the amount of oversampling that is needed. Since it just duplicates existing samples, this might lead to an increase in the overfitting of the classification algorithm (Ling. 1998).

SMOTE: Synthetic Minority Over-Sampling Technique.

This is a synthetic minority over sampling technique. Here, the minority class samples are artificially generated by considering the "feature space" of the dataset and its nearest neighbours. These synthetic samples balance the ratio between the minority and majority classes without changing the majority class. SMOTE first identifies the feature vector and its nearest neighbours and then takes the difference of the distance between them. The number of nearest neighbours can be selected depending on the amount of oversampling required. Then the difference is multiplied by a random number and is identified as a new data point on the line between them. The same procedure follows until both classes are balanced (Chawla 2002).

Table 2: Features of the Data Set

Feature	Data Type	Data Range
High Cholesterol	Binary	0 = No high cholesterol 1 = High cholesterol
Cholesterol Check	Binary	0 = Hasn't checked cholesterol within past five years 1 = Has checked cholesterol within past five years
BMI	Numeric	[12,98]
Smoking	Binary	0 = Has smoked more than 100 cigarettes. 1 = Hasn't smoked more than 100 cigarettes
Stroke	Binary	0 = Hasn't suffered from a Heart Stroke 1 = Has suffered from a Heart Stroke
Diabetes	Numeric	0 = No Diabetes 1 = Diabetes 2 = Only during the pregnancy
Physical Activity	Binary	0 = Hasn't exercised during the past 30 days 1 = Has exercised during the past 30 days
Fruit	Binary	0 = Hasn't consumed at least 1 fruit per day 1 = Has consumed at least 1 fruit per day
Vegetable	Binary	0 = Hasn't consumed at least 1 vegetable per day 1 = Has consumed at least 1 vegetable per day
Alcohol Consumption	Binary	0 = Men: Less than 14 drinks per week Women: Less than 7 drinks per week 1 = Men: more than 14 drinks per week Women: more than 7 drinks per week
Health Care is covered by a Health Insurance.	Binary	0 = No 1 = Yes
Did not meet a doctor during the past 12 months due to financial issues	Binary	0 = No 1 = Yes
General Health Rating	Numeric	[0,5]
Mental Health Rating	Numeric	[0,30]
Physical Health Rating	Numeric	[0,30]
Difficulties in walking	Binary	0 = No 1 = Yes
Sex	Binary	0 = Female 1 = Male
Age	Numeric	[1,6]
Education	Numeric	[0,30]
Income	Numeric	[1,8]

ADASYN: Adaptive Synthetic Sampling.

ADASYN is also a nearest neighbour based algorithm which is similar to the SMOTE algorithm.
The main difference between them is the ADASYN focuses more on the minority data samples which are harder to learn rather than easier to learn data samples. And also, in the SMOTE, it just picks new data points along the straight lines between neighbours. But the ADASYN algorithm looks deeper into the nearest neighbour region by considering the majority class data points inside the region. ADASYN generate synthetic samples only if there are majority samples

inside the neighbour region (Bai 2008).

Under Sampling.

Under sampling techniques focus on the majority class and try to balance both classes by eliminating samples from the majority class. But this might lead to losing important data about the dataset. Hence, this causes to reduce the performance of the machine learning models (Kotsiantis, 2006). If the ratio between minority and majority classes is high, this can cause a lack of data for the analysis. Since these methods drop samples from the majority class, the randomness of the dataset no longer exists. Tthus, the representation of the original target distribution cannot be expected by that the sample.

RUS: Random Under Sampling.

In this under sampling technique, the class imbalance problem is solved by removing samples randomly from the majority data set until two sets are balanced. This can lead to loss of valuable information about the dataset hence reducing the accuracy of the predictions (Yen S.J. 2006).

TOMEK: Tomek Link Under Sampling

Tomek links (TomeK I. 1976) can be defined as follows: given two examples Ei and Ej belonging to different classes, and d(Ei, Ej) is the distance between Ei and Ej . This pair is called a Tomek link if there is not an example El, such as that d(Ei, El) < d(Ei, Ej) or d(Ej , El) < d(Ei, Ej). These Tomek Links can be used as a Under Sampling method to remove the majority class samples to balance the data set. If we want to use this approach as a data cleaning method, we can remove both samples from the majority and minority classes if the above condition for Tomek Link is satisfied.

Cluster Centroid Under Sampling

As per Cluster Centroid Under Sampling method, first the whole data set is divided into some distinct clusters using k means clustering algorithm. Then those clusters are classified by considering the ratio between minority class samples and majority class samples. Initially, the number of cluster samples was set equal to the number of samples in the minority class. Then by using the k mean algorithm, cluster centroids are calculated over the majority class. The calculated cluster centroids are used to replace the entire majority class. This procedure continues until both classes are balanced (Yen , 2009).

Hybrid Sampling

Hybrid Sampling is a combination of the Over Sampling and Under Sampling techniques. Over Sampling increases the data size by adding synthetic information to the minority class while Under Sampling removes data points from the majority class causing a loss of information. A portion of the sampling in the hybrid approach is carried out using oversampling techniques, and the remaining piece is carried out using undersampling techniques. This approach leads to improving the strengths of each technique by reducing the drawbacks (Seiffert, 2009). Many studies have shown that this approach improves the overall performance of the classification algorithm drastically.

SMOTETomek: SMOTE + Tomek Link

This hybrid approach combines SMOTE as an over sampling method and Tomek Link as an Under Sampling method (Wang Z.H.E. 2019). First, the SMOTE sampling is used to generate a new synthetic sample set. Then the newly created data set is processed with Tomek Link to remove Tomek Link pairs from the dataset. The resulting dataset is a balanced dataset with a reduced overlapping between data points.

SMOTEEN: SMOTE + ENN

This is also a hybrid version of the SMOTE over sampling method and the Edited Nearest Neighbour (ENN) under sampling method. In this method, the minority class and the majority class are balanced by using SMOTE technique. Then the balanced dataset is applied to ENN under the sampling process as a cleaning mechanism to eliminate noises generated by the SMOTE while introducing new synthetic samples (Srivastava, 2022).

Data Classification

Applying different Sampling methods discussed in the above section on the BRFSS dataset, we obtain

different balanced datasets for each sampling method. All the generated sample sets are divided into two portions. 70% for training and 30% for testing. Then we employed the following supervised machine learning algorithms to evaluate and compare the performance of each sampling technique.

1. K Nearest Neighbour (KNN) Algorithm.
2. Gaussian Naïve Bayes (Gaussian NB) Algorithm.
3. Decision Tree (DT) Algorithm.
4. eXtreme Gradient Boosting (XGBoost) Algorithm.
5. Light Gradient Boosting Machine (LGB) Algorithm.
6. Adaptive Boosting (ADABoost) Algorithm.
7. Random Forest (RF) Algorithm.

4. RESULTS

In order to analyze the results of our study and to compare the accuracies we will use the confusion matrix and also metrics such as precision, recall, and f1-score.

We will analyze the performance of each machine learning algorithm by evaluating the following statistical parameters.

Confusion Matrix.

The confusion matrix has four important parameters to summarize the performance of the machine learning classifier.

1. TP (True Positive): The total number of data samples where the model correctly predicts the positive class.
2. TN (True Negative): The total number of data samples where the model correctly predicts the negative class
3. FP (False Positive): The total number of data samples where the model incorrectly predicts the negative class

4. FN (False negative): The total number of data samples where the model incorrectly predicts the positive class

Accuracy.

This represents the baseline performance of the classification model. This is calculated by taking the ratio between correctly predicted classes (TP + TN) and the total number of samples.

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP + TN}}{\boldsymbol{TP + TN + FP + FN}}$$

Precision.

This is refered to as the false positive rate. This is calculated by taking the ratio between correctly predicted positive class (TP) and the total number of positive predictions (TP + FP).

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FP}}$$

Recall:

Recall or Sensitivity is calculated by taking the ratio between correctly predicted positive class (TP) and the total number of actual positive samples.

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FN}}$$

Mathews Correlation Coefficient (MCC)

This is a balanced method of all the parameters of the confusion metrics. This can be used even when the class sizes are not equal.

$$\boldsymbol{MCC} = \frac{\boldsymbol{TP * TN - FP * FN}}{\sqrt{(\boldsymbol{TP + FP})(\boldsymbol{TP + FN})(\boldsymbol{TN + FP})(\boldsymbol{TN + FN})}}$$

F1 Score:

This is a mean indicator of the precision and recall of the dataset. This is calculated by taking the harmonic mean of both parameters.

$$\boldsymbol{F1\ Score} = \frac{\boldsymbol{2 * Recall * Precision}}{\boldsymbol{Recall + Precison}}$$

The preprocessed dataset was applied to the above machine learning algorithms in Python version 3.8. The following libraries were utilized in the classification process.

- Pandas version 1.4.4.
- Sklearn version 1.1.2.
- Seaborn version 0.12.
- Matplot version 3.5.3.
- Imbalanced learn 0.9.1.

By applying all the sampling methods mentioned earlier in the dataset with above mentioned machine learning algorithms, we calculated the accuracies, recall values, MCC values, and F1 values for each case and presented them in the following tables:

Table 3: Confusion Matrix

		Predicted Value	
		Has/had a heart disease	Doesn't have/ Didn't have a heart disease
Actual Value	Has/had a heart disease	**True Positive (TP)**	**False Negative (FN)**
	Doesn't have/ Didn't a heart disease	**False Positive (FP)**	**True Negative (TN)**

Table 4: Accuracy values for Sampling methods.

Algorithm	**SMOTE**	**ADSYN**	**ROS**	**RUS**	**TOMEK**	**CLUSTER**	**SMOTEK**	**SMETEEN**
KNN	0.8573	0.8524	0.8772	0.7208	0.8970	0.6300	0.8595	**0.9543**
GNB	0.7512	0.7437	0.7279	0.7244	**0.8240**	0.8000	0.7513	0.8128
DT	0.9155	0.9166	**0.9497**	0.6737	0.8532	0.9349	0.9177	0.9441
XGBOOST	0.9477	0.9478	0.7969	0.7661	0.9052	**0.9669**	0.9487	0.9607
LGBOOST	0.9436	0.9446	0.7780	0.7700	0.9072	**0.9685**	0.9445	0.9576
ADABOOST	0.9050	0.9046	0.7698	0.7678	0.9091	**0.9601**	0.9070	0.9389
RF	0.9466	0.9467	**0.9710**	0.7607	0.9022	0.9623	0.9476	0.9650

Table 5: Precision values for Sampling methods.

Algorithm	**SMOTE**	**ADSYN**	**ROS**	**RUS**	**TOMEK**	**CLUSTER**	**SMOTEK**	**SMETEEN**
KNN	0.8853	0.8826	0.8986	0.7210	0.8655	0.6440	0.8861	**0.9572**
GNB	0.7521	0.7444	0.7288	0.7249	**0.8834**	0.8019	0.7520	0.8197
DT	0.9156	0.9167	**0.9538**	0.6738	0.8597	0.9349	0.9177	0.9440
XGBOOST	0.9507	0.9510	0.8001	0.7678	0.8790	**0.9672**	0.9515	0.9610
LGBOOST	0.9466	0.9479	0.7811	0.7748	0.8828	**0.9687**	0.9473	0.9580
ADABOOST	0.9053	0.9053	0.7703	0.7678	0.8857	**0.9609**	0.9072	0.9389
RF	0.9484	0.9487	**0.9724**	0.7625	0.8743	0.9632	0.9476	0.9652

When considering Table 3 accuracy results, we can see that the Random Forest algorithm has obtained an accuracy higher than 0.9000 for all sampling methods except Random Under Sampling Method. SmoteTomek, Smoteen hybrid random sampling methods, and the Cluster Centroid method have shown an accuracy greater than 0.9000 for Decision Tree, XGBoost, LGBoost, ADABoost, and Random Forest algorithms. These hybrid algorithms have shown an accuracy above 0.75 for all the classification algorithms employed in the study.

As shown in Table 4,5,6,7 XGBoost, LGBoost, and ADABoost algorithms have shown above 0.92 values for precision, recall, MCC score, and F1 score for the cluster centroid algorithm. And also, SmoteTomek and Smoteen hybrid algorithms have shown values

above 0.90 for precision, recall, MCC, except F1 score, which is also above 0.80. TomekLink algorithm has shown the poorest performance in terms of the MCC value for all the machine learning algorithms as shown in table 6.

Table 6: Recall values for Sampling methods.

Algorithm	SMOTE	ADSYN	ROS	RUS	TOMEK	CLUSTER	SMOTEK	SMETEEN
KNN	0.8573	0.8524	0.8772	0.7208	0.8970	0.6300	0.8595	**0.9543**
GNB	0.7512	0.7437	0.7279	0.7244	**0.8240**	0.8000	0.7513	0.8128
DT	0.9155	0.9166	**0.9497**	0.6737	0.8532	0.9349	0.9177	0.9441
XGBOOST	0.9477	0.9478	0.7969	0.7661	0.9052	**0.9669**	0.9487	0.9607
LGBOOST	0.9436	0.9446	0.7780	0.7700	0.9072	**0.9685**	0.9445	0.9576
ADABOOST	0.9050	0.9046	0.7697	0.7678	0.9091	**0.9601**	0.9070	0.9389
RF	0.9466	0.9467	**0.9710**	0.7607	0.9022	0.9623	0.9476	0.9650

Table 7: MCC values for Sampling methods.

Algorithm	SMOTE	ADSYN	ROS	RUS	TOMEK	CLUSTER	SMOTEK	SMETEEN
KNN	0.7420	0.7343	0.7756	0.4410	0.2002	0.2748	0.7450	**0.9079**
GNB	0.5033	0.4880	0.4566	0.4492	0.3058	0.6020	0.5034	**0.6240**
DT	0.8312	0.8332	**0.9035**	0.3475	0.1985	0.8698	0.8354	0.8842
XGBOOST	0.8985	0.8988	0.5970	0.5340	0.2512	**0.9341**	0.9001	0.9193
LGBOOST	0.8902	0.8925	0.5592	0.7700	0.2444	**0.9372**	0.8917	0.9129
ADABOOST	0.8102	0.8099	0.5402	0.5362	0.2777	**0.9210**	0.8142	0.8738
RF	0.8950	0.8954	0.9434	0.5232	0.2401	0.9255	0.8970	**0.9650**

Table 8: F1 values for Sampling methods

Algorithm	SMOTE	ADSYN	ROS	RUS	TOMEK	CLUSTER	SMOTEK	SMETEEN
KNN	0.8547	0.8494	0.8756	0.7204	0.8741	0.6222	0.8570	**0.9539**
GNB	0.7511	0.7434	0.7275	0.7242	**0.8475**	0.7997	0.7512	0.8140
DT	0.9155	0.9166	**0.9496**	0.6736	0.8564	0.9349	0.9177	0.9440
XGBOOST	0.9476	0.9478	0.7963	0.7658	0.8806	**0.9669**	0.9486	0.9607
LGBOOST	0.9435	0.9445	0.7774	0.7693	0.8792	**0.9685**	0.9445	0.9577
ADABOOST	0.9049	0.9046	0.7697	0.7677	0.8869	**0.9601**	0.9070	0.9389
RF	0.9465	0.9467	**0.9710**	0.7603	0.8775	0.9623	0.9475	0.9650

Receiver Operating Characteristic (ROC Curve) is another common technique to compare the performance of the classification algorithms. A ROC curve represents a trade off between the true positive rate and the false positive rate.

Following figures show the ROC curves for each sampling method for all 7 machine learning algorithms.

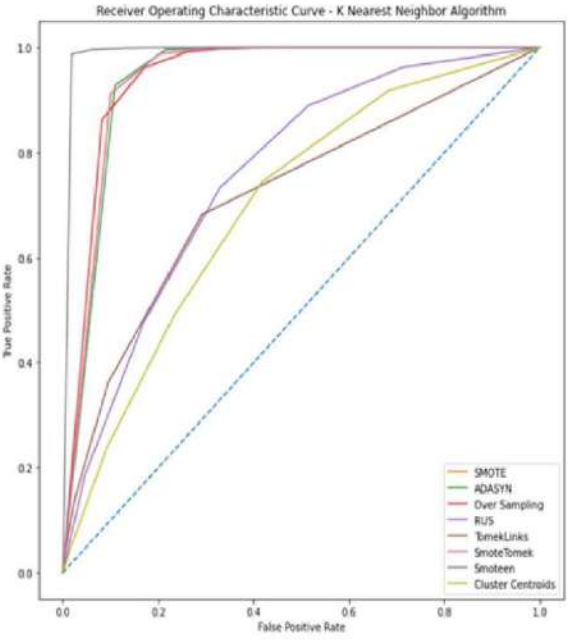

Figure 1: ROC Curves KNN

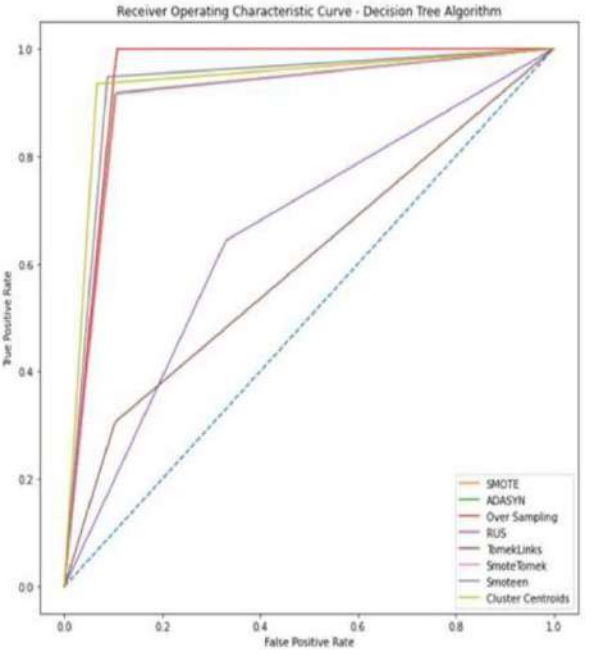

Figure 2: ROC Curves Decision Tree

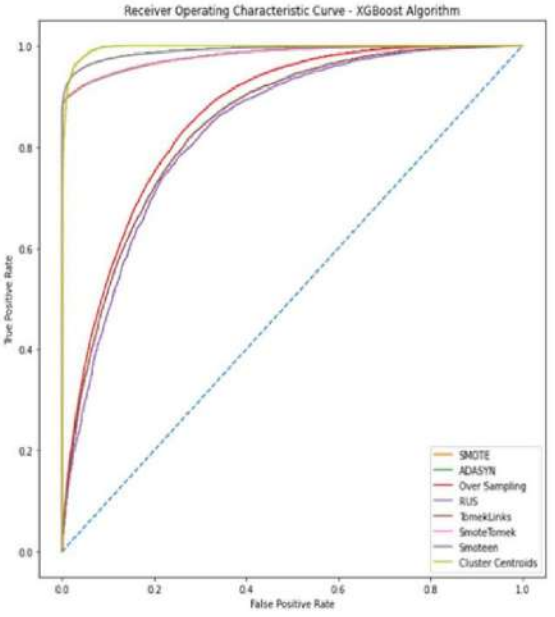

Figure 3: ROC Curves XGBoost

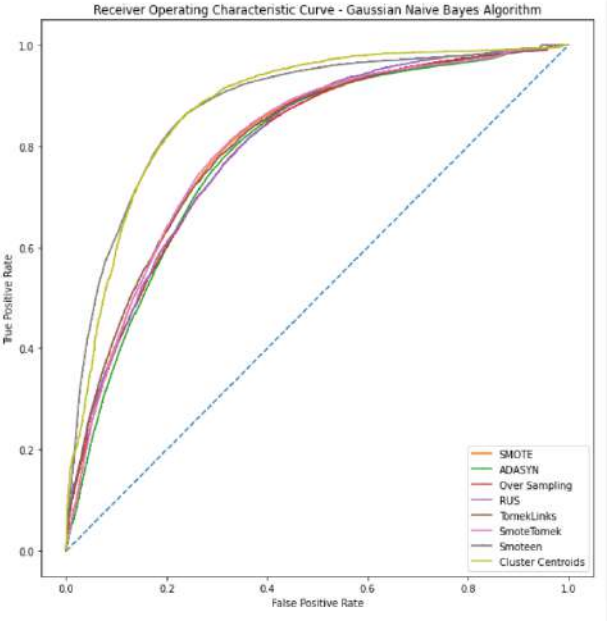

Figure 4: ROC Curves Gaussian Naïve Bayes

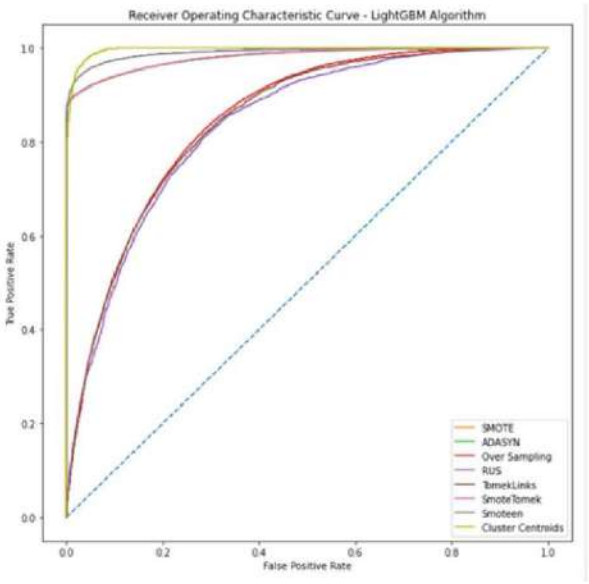

Figure 5: ROC Curves LGBoost

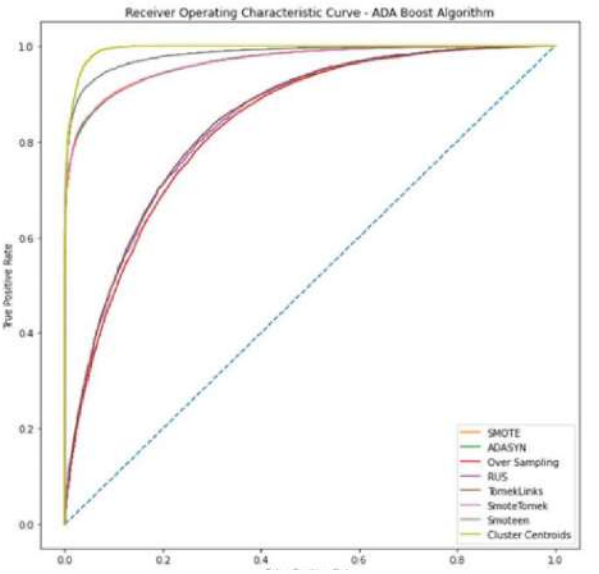

Figure 6: ROC Curves ADABoost

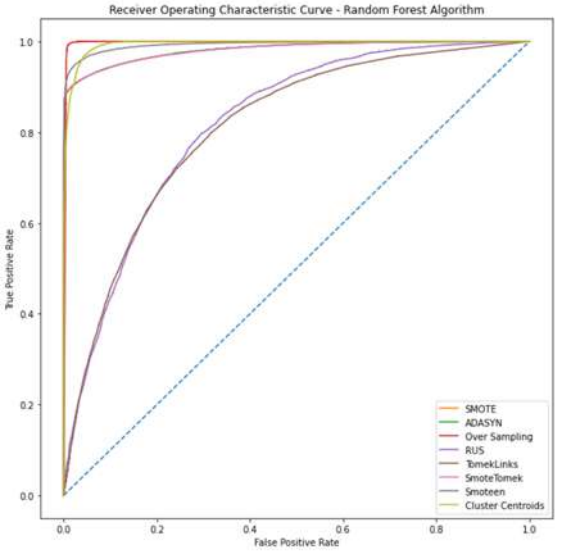

Figure 7: ROC Curves Random Forest

5. DISCUSSION

In this study, we observed that the Random Forest algorithm has shown accuracy above 90% for all the sampling techniques except the Random Under Sampling method. This is because it is an ensemble method, and performed very well on every class imbalance method. All the hybrid methods SmoteTomek and Smoteen show the best performance for all classification algorithms specially for Decision Tree, XGBoost, LGBoost, ADABoost, and Random Forest algorithms. These hybrid methods show above 90% accuracy for all the machine learning algorithms except for k nearest neighbor and Gaussian Naive

Bayes algorithms. This is not only for accuracy, but also valid for other measured statistical parameters as well. From the analysis of all the class imbalance methods, it is evident that Over Sampling followed by Under Sampling methods can improve the performance of the classifier drastically because of its behavior. In the presented work, heart disease detection compares eight class imbalance methods over seven classifiers.

6. CONCLUSION

The performance of ensemble classifiers AdaBoost, Random Forest, and XGBoost is better than the base classifiers mainly due to their ensemble behaviour; somehow, KNN and Decision Tree classifier also performed very well. In all base classifiers, the performance of the Gaussian Naive Bayes classifier was the least in each class imbalance method. In this study we have shown that in heart disease prediction not only the classification algorithm but also the sampling techniques are important when dealing with a class imbalanced dataset. Because if you have an imbalanced dataset it will result in very low accuracies for the respective classification algorithms and as a result will not be able to detect heart diseases accurately. We also showed that for validation, only considering the accuracy metric is not sufficient.

7. REFERENCES

Ramesh, T.R., Lilhore, U.K., Poongodi, M., Simaiya, S., Kaur, A. and Hamdi, M., (2022). Predictive Analysis of Heart Diseases With Machine Learning Approaches. *Malaysian J of Computer Science*, pp.132-148.

Palaniappan, S. and Awang, R., 2008, March. Intelligent heart disease prediction system using data mining techniques. In *2008 IEEE/ACS int conf on computer systems and applications.* pp. 108-115.

Rajdhan, A., Agarwal, A., Sai, M., Ravi, D. and Ghuli, P.,(2020). Heart disease prediction using machine learning. *International Journal of Research and Technology*, *9*(04), pp. 659-662.

Bahrami, B. and Shirvani, M.H., 2015. Prediction and diagnosis of heart disease by data mining techniques. *Journal of Multidisciplinary Engineering Science and Technology (JMEST)*, *2*(2), pp.164-168.

Tadiparthi, P.K. and Kuna, V., 2022. Heart Disease Prediction Using Machine Learning Algorithms: A Systematic Survey.

Li, J.P., Haq, A.U., Din, S.U., Khan, J., Khan, A. and Saboor, A., (2020). Heart disease identification method using machine learning classification in e-healthcare. *IEEE Access*, *8*, pp. 107562-107582.

Ishaq, A., Sadiq, S., Umer, M., Ullah, S., Mirjalili, S., Rupapara, V. and Nappi, M., 2021. Improving the prediction of heart failure patients' survival using SMOTE and effective data mining techniques. *IEEE access*, *9*, pp. 39707-39716.

Fitriyani, N.L., Syafrudin, M., Alfian, G. and Rhee, J., 2020. HDPM: an effective heart disease prediction model for a clinical decision support system. *IEEE Access*, *8*, pp.133034-133050.

Snead, R., Dumenci, L. and Jones, R.M., 2022. A latent class analysis of cognitive decline in US adults, BRFSS 2015-2020. *BMC Public Health*, *22*(1), pp.1-10.

M. Galar, A. Fernandez, E. Barrenechea, H. Bustince and F. Herrera, "A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches," in *IEEE Transs on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4), pp. 463-484, Jul 2012, doi: 10.1109/TSMCC.2011.2161285.

Singh, A. and Purohit, A., 2015. A survey on methods for solving data imbalance problem for classification. *Int J of Computer Applications*, *127*(15), pp.37-41.

Rančić, S., Radovanović, S. and Delibašić, B., 2021, May. Investigating oversampling techniques for fair machine learning models. In *Int Conf on Decision Support System Technology*. pp. 110-123. Springer, Cham.

Kovács, G., 2019. An empirical comparison and

evaluation of minority oversampling techniques on a large number of imbalanced datasets. *Applied Soft Computing*, *83*, p.105662.

Ling, C.X. and Li, C., 1998, August. Data mining for direct marketing: Problems and solutions. In *Kdd* 98, pp. 73-79.

Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. *J of artificial intelligence research*, *16*, pp.321-357.

He, H., Bai, Y., Garcia, E.A. and Li, S., 2008, June. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE int joint conf on neural networks (IEEE world congress on computational intelligence)*.pp. 1322-1328)..

Kotsiantis, S., Kanellopoulos, D. and Pintelas, P., 2006. Handling imbalanced datasets: A review. *GESTS int trans on computer science and engineering*, *30*(1), pp. 25-36.

Yen, S.J. and Lee, Y.S., 2006. Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset. In *Intelligent Control and Automation*.pp. 731-740. Springer, Berlin, Heidelberg.

Tomek, I., 1976. An Experiment with The Edited Nearest-Nieghbor Rule.

Yen, S.J. and Lee, Y.S., 2009. Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Systems with Applications*, *36*(3), pp. 5718-5727.

Seiffert, C., Khoshgoftaar, T.M. and Van Hulse, J., 2009. Hybrid sampling for imbalanced data. *Integrated Computer-Aided Engineering*, *16*(3), pp. 193-210.

Wang, Z.H.E., Wu, C., Zheng, K., Niu, X. and Wang, X., 2019. SMOTETomek-based resampling for personality recognition. *IEEE Access*, *7*, pp.129678-129689.

Srivastava, J. and Sharan, A., 2022. SMOTEEN Hybrid Sampling Based Improved Phishing Website Detection.