2022 Second International Conference on Computer Science, Engineering and Applications (ICCSEA) | 978-1-6654-5834-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICCSEA54677.2022.9936470

Improving Prediction of Chronic Heart Failure using SMOTE andMachine Learning

[1]Sudipta Priyadarshinee,[2]Madhumita Panda

[1]Department of Computer Science, Research Scholar, G.M. University, Sambalpur, Odisha, India

sudiptapatel88@gmail.com[2]Department of Computer Science, Associate ProfessorG.M. University, Sambalpur, Odisha, India

mpanda.gmu@gmail.com

Abstract- Around the world, cardiovascular disease (CVD) is a vital reason for death and morbidity.Predicting heart disease survivors is a significant challenge in clinical data analytics.Machine learning converts massive volumes of raw data created by the healthcare business into meaningful knowledge that can aid in decision-making. Several research have shown that important attributes are important in increasing the accuracy of machine learning approaches. This paper looks at heart failure survivors from a group of 299 individuals who were hospitalised to the hospital. The goal is to use of machine learning models that can enhance the predictability of cardiac patient survival. Thispaper employs eight classification models: Random Forest (RM), Extra Tree (ET), NaïveBayes (NB), K Nearest Neighbour (KNN), Decision Tree J48, Decision Table/Naïve Bayes hybrid classifier (DTNB), Optimized Forest, Alternating Decision Tree (ADTree) to predict patient's survival. Synthetic Minority Oversampling Technique (SMOTE) is used to resolve the unbalance dataset. Experiment outcomes demonstrate SMOTE technique improves the accuracy of the selected classifier's output and DTNB achieves highest accuracywith 87.08%utilising SMOTE to forecast the survival of cardiac patients. All experiments are carried out in a simulation environment using the WEKA tool.

Keywords-Cardiovascular Disease, Machine Learning, WEKA, SMOTE

I. INTRODUCTION

The most prevalent reason for death is heart disease in theglobe, based on the World Health Organization (WHO) [1].Because of a few contributing factors that contribute to cardiovascular disease, such as cholesterol levels,high blood pressure, incorrect pulse rate,diabetics and a variety of additional factors, identifying cardio vascular disease (CVD) is fairly challenging [2].The process of detecting or predicting heart disease from patient records is known as heart disease diagnosis.Diagnosing heart disease is a difficult task that necessitates experience and knowledge. An incorrect diagnosis may result in the patient's death or disability. The Disease Prediction Model can help medical professionals and practitioners predict heart disease. The enormous volume of information that can be acquired with the use of digital gadgets (either by the patient or in the hospital) can be combined approaches of machine learning to identify and forecast diseases.

One among the most popular quickly expanding sector of artificial intelligence is machine learning.These algorithms can analyse vast amounts of information from a various sector, such as the medical field.By lowering the error in prediction and actual results, a number of machine learning methods are utilised to better grasp interaction between multiple components that is complicated and non-linear[3].Algorithms of Machinelearning must be used to aid medical practitioners in analysing data and producing diagnostic decisions that are exact and accurate, because of the ever-increasing volume of medical information. Different classification methods are employed in medical data mining to forecastcardiovascular disease in patients and death forecasts owing to heart attacks [4].

The remainder of the document is arrangedin the following fashion:The summery of literature on various heart-related works is highlighted in Segment II. The proposed framework, as well as the several algorithms utilised to classify the given dataset, are described in Segment III. Segment IV explains the outcomes of implementing the proposed methodology. The conclusions of the suggested effort are finally summarised in Segment V.

II.RELATED WORK

In this paper [5] the authors employed eight algorithms: Decision Tree, Logistic model tree algorithm,J48 algorithmSupport Vector Machine, Nave Bayes, RandomForest andKNNfor predicting the onset of cardiac disease. According to the findings of the experiments, it is found that the J48 tree technique is shown to be the best algorithm for predicting cardiac disease since it has the good accuracy and requires the shortest amount of time to construct.

978-1-6654-5834-4/22/$31.00 ©2022 IEEE

The authorssuggested[6] a system for studying various heart conditions and primary causes of death. Various algorithms of machine learning likeDecision Tree ,Nave Bayes, Random Forest, and KNN were used. Only 14 of the 76 attributes were used because their research goal is to create a precise and effective system with smaller number of features KNN outperformed the other three supervised machine learning classifiers.

In this paper [7] the clinical support system (CDSS) was investigated for cardiacfailure analysis. In their study,they contrasted the performance of several machine learning algorithm. With a 87.6 percent accuracy, the CART provided the best outcome.

The authors in [8] concentrated on a diabetic patient with heart problems. They used a variety of predictors, including blood pressure, blood sugar, and age. Using the SVM classifier, they managed to achieve a 94.60 percent accuracy rate. The data set was unbalanced, and the authors failed to address this problem.

Utilizing SMOTE (Synthetic Minority Oversampling Technique) on all the nine classification approaches [9] used to predict patient survival, the datasets were equalized. The highest-rated features chosen by Random Forest was used to train machine learning systems. In predicting the survival of people with a cardiac issue, Extra Tree Classifier has been proven to perform well in experiments, with an accuracy of 0.9262 using SMOTE.

In this paper [10] the authors presented a low-cost solution for predicting heart attacks with high level of accuracy and reliability. It predicts heart attacks using several types of machine learning classifiers on a UCI dataset without the use of feature engineering. To managethe imbalance data, the proposed project employs a SMOTE (Synthetic Minority Oversampling Technique). From experimental result it is found that SMOTE-based artificial neural networks performed the best amongst all machine learning classifiers.

The term "crime type" is used in this paper [11] to describe how law-breaking individuals are classified. Exploration measured an improvement in the wrongdoing expectation model using the Rule Based Decision Tree (RBDT J48) calculation and Naive Bayes, as it is the most effective AI calculation for predicting wrongdoing information.

A neuroevolution system has been created for forecasting a number of plant diseases in this study [12]. The disease's prediction is based on weather parameters that are linked to climate change data, as well as the soil in the area. The authors demonstrated how to use an ANN-based neuroevolution model to predict several plant diseases in this paper.

Several Machine Learning researchers provided numerous algorithms and methodologies for detecting phishing websites in this publication [13]. The majority of the work was completed using well-knownmethodsthey are Naive Bayesian, SVM, Random Forest and Decision Tree, according to the author. For detection, some authors proposed new systems like PhishScore and PhishChecker.

The author used five different machine learning techniques [14] to make predictions about cardiac disease: the Support Vector Machine,Logistic Regression,Decision Tree K-Nearest Neighbor, and Random Forest. The Random Forest machine learning classifier, in contrast, runs in 1.09 seconds and achieves an optimal accuracy of 85% based on its ROC AUC score of 0.8675.

Several different types of supervised machine learning methods were employed to categorize a cardiovascular dataset in the research [15]. Among these methods were Logistic Regression,Naive Bayes,Random Forest, Decision Tree, KNN, and SVM. With a 73 percent accuracy, the Decision Tree provided the best outcome.

III.PROPOSED FRAMEWORK

In this researchwe haveapplied eight classification models: Random Forest (RM), Extra Tree (ET), Naïve bayes (NB), K Nearest Neighbour (KNN) Decision Tree J48, Decision Table/Naïve Bayes hybrid classifier (DTNB), Optimized Forest, Alternating decision tree (ADTree) to predict patient's survival.To address the data imbalance problem, the SMOTE approach is used to the given dataset. The algorithms of machine learning are thenemployedon the balanceddataset and the rate ofaccuracy is calculated. The primary goal was to find the method that could best classify the given dataset.

A.Dataset

The Kaggle [27] heart dataset has been used for the experiment. The dataset includes 299 patients' medical records with heart problems who were accumulatedduring the time of follow-up, with each profile of the patient containing 13 clinical attributes.There are 194 men and 105 women among the 299 records. All of the patients are beyond the age of 40. In the target class, 1 denotes the deceased and 0 denotes the alive. The Table 1 provides an overview of the data set.

Table 1. Specification Of the Dataset

Sr No	Attributes	Description	Measured In	Range
1	Age	The patient's age	Years	40 - 95
2	Anaemia	Red blood cell or haemoglobin deficiency	Boolean	0, 1
3	Creatinine phosphokinase (CPK)	CPK enzyme levels in the blood	mcg/L	23 -7861
4	Diabetes	If the patient suffers from diabetes	Boolean	0, 1
5	ejection fraction	% Of blood leaving	Percentage	14 - 80
6	high_blood_pressure	If a patient has hypertension	Boolean	0, 1
7	Platelets	The number of platelets in the blood	kilo platelets/Ml	25.01-850.00
8	serum creatinine	Creatinine concentration in the blood	mg/dL	0.50 -9.40
9	serum sodium	Sodium levels in the blood	mEq/L	114 -148
10	Sex	Woman or man	Binary	0, 1
11	Smoking	If the patient is a smoker	Boolean	0, 1
12	Time	Period of follow-up	Days	4 - 285
13	(Target)death event	If the patient died during the period of follow-up,	Boolean	0, 1

C. Imbalance Nature of Dataset

In total, there are 203 instances of the negative class (0) and 96 of the positive class (1).Within the dataset, there is an unbalanced distribution of classes. One of the key causes of decreased classification model accuracy is unequal distribution. As a result of their imbalance problem in a dataset, most machine learning algorithms are unable to discover patterns for both positive and negative classes well. Furthermore, as the minority class, that is the positive class, is few in number, the outcomes provided by this class are typically unsuccessful. The imbalanced characteristics of the presented dataset is well tackled by SMOTE approach, which is one of the proposed work's significant contributions. Fig. 1 shows the original heart data set in WEKA software.

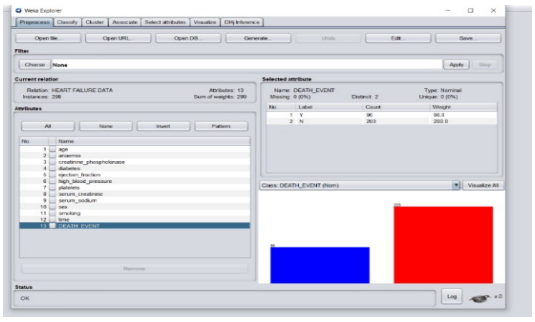

Fig. 1. Original heart data set in WEKA software.

D. SMOTE (Synthetic Minority Oversampling Technique)

A technique for oversampling is SMOTE methodology. that is frequently utilised in medicine to cope with data that is class unbalanced [16]. When dealing with imbalanced datasets, the SMOTE pre-processing technique is considered one of the most dependable and powerful pre-processing strategies in the machine learning and information mining industry[10]. To expand the amount of data instances, Using Euclidean distance, SMOTE generates random false minority data from its closest neighbours.Because new instances are formed based on original features, they become identical to the original data [17]. When working with data that has many dimensions, SMOTE is not the greatest option because it can add to the noise.In this study, after SMOTE technique is used in the dataset,SMOTE raised the number of data samples from 299 to 395 and the dataset is balanced. Fig. 2 shows the Balance heart data set using SMOTE in WEKA software.

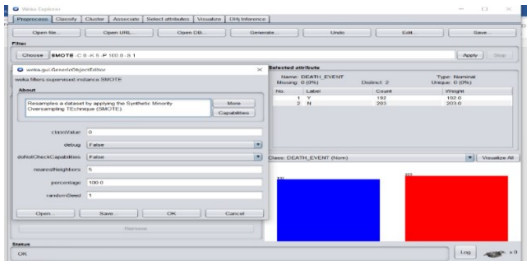

Fig. 2. Balance heart data set using SMOTE in WEKA software.

E.10-folds Cross Validation

Cross validation is a technique of estimating machine learning classifier's performance. It aids researchers in estimating the accuracy of model predictions in practise. In the datasets, there are two types of phases: training sets and testing sets. Cross validation will be used to compare testing and training sets in order to rule out overfitting and identify how the machine learning techniques should produce independent data. [18].

F. Tools and Technique

Weka is useful tool which was utilised to carry out all of the experiments on the classifiers described in this paper. The Weka tool is a gathering of methods of machine learning.It is employed to categorise datasets in an automated fashion using the specified algorithm, for so long as that algorithm is available in the environment.

G.Classification Algorithms

1. **Naive Bayes:**The Bayes' Theorem is utilized to the construct aclass of classification algorithms called as Naive Bayes classifiers. It is aset of algorithms, not a single algorithm, that all share a basic principle.That is, each pair of classification features is different from the others [19].
2. **Random Forest:** As a final output, the random forest (RF) algorithm creates a number of decision trees and finds the mode of all classes output by each tree. There may be high variance in a single tree, but RF alleviates this problem [20].
3. **K – Nearest Neighbour:**One of the simplest yet most effective categorization techniques is the K-Nearest Neighbour (KNN) method. The first k data points in the training set that are most closely connected to the data point for which a target value isn't available are found using this methodand assigns the average value of the data points identified to it [21].
4. **Extra Tree:** An ensemble machine learning algorithm is Extra Tree. It's an ensemble of decision tree methods that works in conjunction with others, such bootstrap aggregation and random forest. It is conceptually related to the Random Forest approach [22].
5. **Decision Tree J48:** J48 is an upgrade to C4.5 that uses a decision tree-based method. With this procedure, a tree is built to demonstrate the grouping procedure in decision tree. The interior nodes of the tree represent a test on an attribute, the branches show the outcome of the test, the class mark is held by a leaf node and the topmost node is the root node. The output of the decision tree predicts new cases of data [23].
6. **Decision Table/Naïve Bayes hybrid classifier (DTNB):**It is for creating and deploying a hybrid decision table/naive bayes algorithm. At every stage of the search, the classifier weighs the benefits of separating the characteristics into two separatesubclasses, one is decision table and another is naïve bayes [24].
7. **Optimized Forest:**The Forest Optimization (FOA) Algorithm is another method for solving nonlinear problems with optimization, which is inspired by natural processes in forests. Making use of a genetic algorithm to optimise the number of trees in a decision forest in order to find a sub forest with good rate of ensemble accuracy. [25].
8. **AD Tree:**A machine learning classification strategy known as an alternate decision tree (AD Tree). It is related to boosting and generalises decision trees. An AD Tree is made up of prediction nodes that carry a single number and decision nodes that describe a predicate condition. An AD Tree identifies an instance by walking all pathways with true decision nodes and adding any prediction nodes traversed. [26].

IV. RESULTS AND DISCUSSION

On the provided dataset the experiment is carried out and the outcomes are gathered.Each experiment is subjected to 10-fold cross validation to ensure that the results are free of bias.First,on the given dataset, the experiment is carried out using a wide range of machine learning techniques.ThenSMOTE is employed to balance the dataset. Afterwards,on the balanced dataset, machine learning approaches are utilised and estimated theaccuracy.The classification accuracy without SMOTE is presented in Table 2. The classification accuracy with SMOTE is presented in Table 3.

A. Results of Experiment without SMOTE

On the given dataset, the experiment is carried out using a variety of machine learning methods. In this section, we assess the efficacy of all algorithms in terms of correctly classified instances, incorrectly classified instances, and accuracy. Table 2 displays the outcomes.

B. Results of Experiment with SMOTE

SMOTE (Synthetic Minority Oversampling Technique) is a potent solution to the issue of class which is not balancedand have shownsolid results in a variety of fields. To create a balanced dataset, the SMOTE method adds fake data to the minority class. The results of machine learning algorithms utilising the SMOTE method on the provided dataset are summarized in the Table 3.

Table 3. Classification Accuracy With SMOTE

Algorithm	Balanced heart data set	Correctly classified instances	Incorrectly classified instances	Accuracy (%)
	Total instances			
Naive Bayes	395	310	85	78.48
KNN	395	281	114	71.13
Random forest	395	342	53	86.58
Decision tree J48	395	327	68	82.78
Extra tree	395	308	87	77.97
DTNB	395	344	51	87.08
Optimized Forest	395	342	53	86.58
AD Tree	395	340	50	86.07

Table 3 shows that the SMOTE considerably enhances the effectiveness of all classifiers in all assessment matrices.DTNB classifierand SMOTE improved results by 5% when compared to results obtained without the use of SMOTE.DTNB achieved highest results with 87.08% accuracy. Random forest and optimized forest achieved second good results with 86.58% accuracy.

V. CONCLUSION AND FUTURE WORK

The use of machine learning algorithms to process raw health data from the heart will help save the lives of cardiac patients. The mortality rate can be managed

Table 2. Classification Accuracy Without SMOTE

Algorithm	Original heart data set	Correctly classified instances	Incorrectly classified instances	Accuracy (%)
	Total instances			
Naive Bayes	299	228	71	76.25
KNN	299	203	96	67.89
Random forest	299	250	49	83.61
J48	299	245	54	81.93
Extra tree	299	216	83	72.24
DTNB	299	246	53	82.27
Optimized Forest	299	252	47	84.28
AD Tree	299	247	52	82.60

FromTable 2, it is found that the optimized forest classifier performed well, with an accuracy of 84.28%.Random forest is second good classifier with 83.61% accuracy.

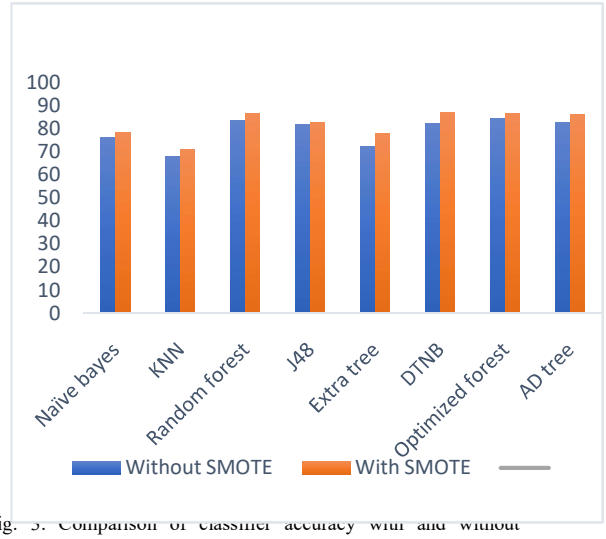

Fig. 3. Comparison of classifier accuracy with and without SMOTE

by evaluating variables that contribute to heart failure and taking preventive actions. This work proposes a machine learning-based strategy that is both effective and efficient for predicting the survival of heart patients. Machine learning methods include Naïve Bayes,K Nearest NeighbourRandom

ForestExtra Tree,Decision Tree J48, DTNB, Optimized Forest, ADTree. SMOTE is used to address the issue of class unbalance. It's also been discovered that using the SMOTE technique improves the accuracy of the selected classifier's output and DTNB achieves highest accuracy with 87.08% with SMOTE in prediction of survival of cardiacpatient. In future we'll apply a variety of classification techniques, particularly ensemble approaches to enhancethe performanceof the models and we will experiment with a several types of approaches to address the unbalanced situation.

REFERENCES

[1] WHO. The Top 10 Causes of Death. Accessed: Dec. 30, 2020. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/the-top-10- causes-of-death

[2] C. Fryar, T.-C. Chen, and X. Li, ''Prevalence of uncontrolled risk factors for cardiovascular disease: United states, 1999-2010,'' in NCHS Data Brief, vol. 103. Aug. 2012, pp. 1–8.

[3] Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machinelearning improve cardiovascular risk prediction using routine clinical data? PLoS ONE. 2017;12(4):e0174944.

[4] Ramalingam VV, Dandapath A, Raja MK. Heart disease prediction using machine learning techniques: a survey. Int J Eng Technol. 2018;7(2.8):684–7.

[5] Kumar, M. Nikhil, K. V. S. Koushik, and K. Deepak. "Prediction of heart diseases using data mining and machine learning algorithms and tools." *International Journal of Scientific Research in Computer Science, Engineering and Information Technology* 3.3 (2018): 887-898.

[6].Shah, Devansh, Samir Patel, and Santosh Kumar Bharti. "Heart disease prediction using machine learning techniques." *SN Computer Science* 1.6 (2020): 1-6.

[7] G. Guidi, M. C. Pettenati, P. Melillo, and E. Iadanza, ''A machine learning system to improve heart failure patient assistance,'' IEEE J. Biomed. Health Informat., vol. 18, no. 6, pp. 1750–1756, Nov. 2014.

[8] G. Parthiban and S. K. Srivatsa, ''Applying machine learning methods in diagnosing heart disease for diabetic patients,'' Int. J. Appl. Inf. Syst., vol. 3, no. 7, pp. 25–30, Aug. 2012

[9] Ishaq, Abid, et al. "Improving the prediction of heart failure patients' survival using SMOTE and effective data mining techniques." *IEEE access* 9 (2021): 39707-39716.

[10] Waqar, Muhammad, et al. "An efficient smote-based deep learning model for heart attack prediction." *Scientific Programming* 2021 (2021).

[11] Lavanya, S., and D. Akila. "Prediction Performance of Crime Against Women Using Rule Based Decision Tree J48 Classification Algorithms in various states of India." *2021 2nd International Conference on Computation, Automation and Knowledge Management (ICCAKM)*. IEEE, 2021.

[12] Kanimozhi, E., and D. Akila. "AN EMPIRICAL STUDY ON MACHINE LEARNING ALGORITHM FOR PLANT DISEASE PREDICTION." *Journal of Critical Reviews* 7.5 (2019): 2020.

[13] Kiruthiga, R., and D. Akila. "Phishing websites detection using machine learning." *International Journal of Recent Technology and Engineering* 8.2 (2019): 111-114.

[14] Kumar, N. Komal, et al. "Analysis and prediction of cardio vascular disease using machine learning classifiers." *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*. IEEE, 2020.

[15] Princy, R. Jane Preetha, et al. "Prediction of cardiac disease using supervised machine learning algorithms." *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*. IEEE, 2020.

[16] R. Blagus and L. Lusa, ''Joint use of over- and under-sampling techniques and cross-validation for the development and assessment of prediction models,'' BMC Bioinf., vol. 16, no. 1, pp. 1–10, Dec. 2015.

[17] N. V. Chawla, ''Data mining for imbalanced datasets: An overview,'' in Data Mining and Knowledge Discovery Handbook. Springer, 2009, pp. 875–886.

[18] Srivastava, H. (2017, December 8). What is K-Fold Cross Validation. Retrieved from https://magoosh.com

[19] Rennie, Jason & Shih, Lawrence & Teevan, Jaime & Karger, David. (2003). Tackling the Poor Assumptions of Naive Bayes Text Classifiers. Proceedings of the Twentieth International Conference on Machine Learning. 41.

[20] L. Breiman, ''Random forests,'' Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.

[21] Mucherino, Antonio, Petraq J. Papajorgji, and Panos M. Pardalos. "K-nearest neighbor classification." *Data mining in agriculture*. Springer, New York, NY, 2009. 83-106.

[22] Sharaff, Aakanksha, and Harshil Gupta. "Extra-tree classifier with metaheuristics approach for email classification." *Advances in computer communication and computational sciences*. Springer, Singapore, 2019. 189-197.

[23] L. Breiman, J. Friedman, R. Olshen, and C. Stone, Classification and Regression Trees (Statistics/Probability Series). 1984.

[24] Sahoo, Soumya, et al. "A hybrid DTNB model for heart disorders prediction." *Advances in electronics, communication and computing*. Springer, Singapore, 2021. 155-163.

[25] Ghaemi, Manizheh, and Mohammad-Reza Feizi-Derakhshi. "Forest optimization algorithm." *Expert Systems with Applications* 41.15 (2014): 6676-6687.

[26] Ooi, Melanie Po-Leen, et al. "Alternating decision trees." *Handbook of Neural Computation*. Academic Press, 2017. 345-371.

[27] https://www.kaggle.com/andrewmvd/heart-failure-clinical-data/discussion/193109