USING MACHINE LEARNING TECHNIQUES FOR PREDICTION OF HEART DISEASE

Dr.Vaka Murali Mohan[1]., Pathikonda Krushika Reddy[2].,P.Varsha Reddy[3].,Chalicheemala Sravani[4].,Nandikunta Anusha[5]

1 Principal &Professor, Department Of CSE., Malla Reddy College Of Engineering For Women., Maisammaguda.,

Medchal.,Ts, India (✉murali_vaka@yahoo.com)

2, 3, 4, 5 B.Tech CSE, (20RG1A05A6, 20RG1A05A5, 21RG5A0509, 20RG1A05A4),

Malla Reddy College Of Engineering For Women., Maisammaguda., Medchal., Ts, India

Abstract

The World Health Organization (WHO) estimates that 12 million deaths worldwide each year are caused by cardiovascular disease. Cardiovascular disorders are the leading cause of mortality worldwide. In high-risk individuals, an early diagnosis of cardiovascular disease may aid in making decisions about modifying their lifestyle. In this study, we apply machine learning methods to the problem of cardiac illness diagnosis. We also used sampling methods for dealing with skewed data. The total risk is predicted using a number of machine learning techniques. Kaggle has a public version of the framingham_heart_disease dataset. Our tests rely on this dataset. The ultimate objective here is to foretell whether or not the patient will develop coronary heart disease (CHD) during the next decade. There are 15 characteristics in the dataset that describe the patients. We improved the identification of cardiac illness by 99% using machine learning methods.

1 INTRODUCTION

Coronary heart disease (CHD) is the most frequent kind of heart disease overall. Every year, it claims the lives of more than 3, 50,000 individuals. About 610,000 Americans succumb to heart disease each year. There are 7.35 million heart attacks annually [5]. Of them, 5.25 million are first-time attacks, while 2.10 million are repeat attacks. Twenty-two percent of all fatalities in Asia are attributable to cardiovascular disease. More than only high blood pressure, diabetes, smoking, high cholesterol, etc., contribute to heart disease. Consequently, cardiac disease is difficult to diagnose. Human heart disease severity has been studied using a variety of data mining and neural network methods. CHD sickness is complex in nature and must be treated with extreme caution because of this. Failure to conduct early detection may have adverse effects on the heart or result in sudden death. Different metabolic disorders are discovered using a medical perspective and data mining. Machine learning is a method that allows a computer to "learn" from examples and data samples without being given any specific instructions. Using previously collected data, machine learning generates new reasoning. The significance of Machine Learning is crucial in many industries. Furthermore,

The study demonstrates its usefulness in the diagnosis of heart disease. Artificial intelligence (AI), which includes Deep Learning, is a kind of machine learning. Many academic disciplines can benefit from using deep learning. It is also used for predicting cardiac issues.

2 LITERATURE SURVEY

Heart disease is the major cause of death and disability worldwide [1]. Machine learning methods were suggested for predicting the risk of cardiovascular disease by Ahmed M. Alaa [2] and colleagues. However, their best accuracy was 77%. The dataset is imbalanced, thus sampling methods must be used. Instead, they used the dataset to test out Machine Learning algorithms. Improvements in cardiovascular risk prediction were the focus of research by Stephen F. Weng [3] and colleagues. Machine learning methods have been demonstrated to increase prediction accuracy for cardiovascular risk, although a larger sample size is needed for optimal results. To better forecast coronary heart disease (CHD), Rine Nakanishi [4, 5] and colleagues examined ML techniques. By using machine learning

techniques, they were able to improve the accuracy of 6814 patient records. Senthilkumar Mohan [6] suggested a machine learning technique to increase the accuracy of cardiovascular illness predictions by identifying key factors. They experimented with several feature combinations until they found that hybrid random forest provided the highest accuracy (88.7%). Good results were found when the K-Nearest Neighbor Algorithm (KNN), Decision Trees (DT), Support Vector Machines (SVM), and Naive Bayes algorithms were used to the problem of predicting cardiovascular illness by Himanshu Sharma [7] et al. Accessibility of datasets has been investigated by Marjia Sultana [8] and colleagues for the visceral, intensely recurrent, and contradicting character of many heart disease illnesses is a hallmark of the condition. These datasets need pre-processing before machine learning methods may be applied to them. In addition, they said that picking the right characteristics is critical for accuracy. A technique for predicting cardiovascular disease was suggested by M.A.Jabbar et al. [9]. They used a genetic algorithm to choose features and then used K-NN, with positive results. A subset of the researchers also used deep learning methods to forecast cardiac events. A deep learning approach with 13 characteristics was suggested by N. Al-milli [10]. If you compare their findings to those of other methods, you'll see that theirs are more accurate.

3 PROPOSED METHOD

We gathered data for cardiovascular disease prediction via Kaggle. A total of 4239 patient records with 15 attributes are included in the dataset. Factors such as age, gender, and health status are examples of features. Factors include chronological age, gender, level of education, current smoking status, daily cigarette consumption, blood pressure medication use, history of stroke or hypertension, diabetes, total cholesterol, blood pressure (both systolic and diastolic), body mass index, heart rate, and blood glucose levels. We need to determine if the patient has a 10-year chance of developing CHD based on these characteristics. Six hundred and forty (644) of the dataset's samples have TenYearCHD = 1, whereas the other samples have TenYearCHD = 0. Machine learning relies heavily on the preceding phase of data pretreatment. Oversampling and under sampling are useful tools for balancing the samples of two classes when working with imbalanced datasets. We used three different sampling methods on the dataset because of its imbalance.

i) Random Over sampling

ii) Synthetic Minority Oversampling

iii) Adaptive synthetic sampling approach

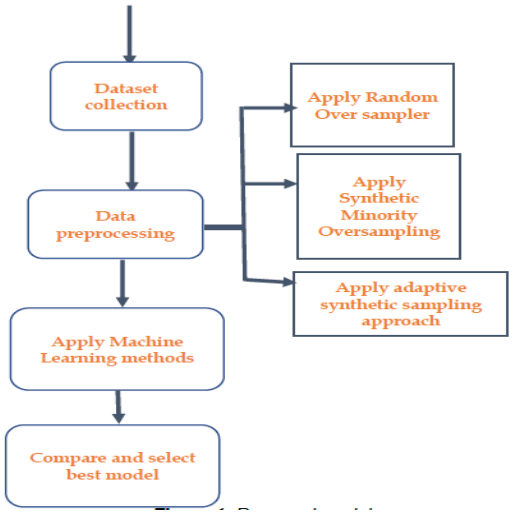

Figure 1: Proposed model

3.1 Random over sampling:

This oversampling method is used to include additional representations of underrepresented groups in the training data. More than one attempt at oversampling should be feasible. This is a fundamental method that has also been shown to be effective. Some examples from the minority group may be picked at random and substituted into the main group. To produce fresh samples from an existing population, a sampling method called "replacement sampling" is used.

3.2 Synthetic Minority Oversampling (SMOTE):

Instead of taking the same samples again and over, SMOTE employs a specialized heuristic approach. Imagine you have n samples of training data and f features in the underlying feature space. Keep in mind these characteristics, think about n samples of training data (with f characteristics).In this method, k represents the number of closest neighbors used to choose a sample from the provided dataset. To generate a synthetic data point, it takes into account a vector between the current data point and one of its k neighbors and multiplies it by a value x between 0 and 1.To make a new point, this pint must be added.

3.3 Adaptive synthetic sampling approach (ADASYN):

Equally reliant on heuristics is ADASYN. It's an SMOTE-based system. The dividing line between easier and harder categories moves. Samples from underrepresented groups are distributed proportionally more heavily to account for their higher average learning difficulty. In contrast to ADASYN, SMOTE does not differentiate between data samples that are easy and difficult to identify using the nearest-neighbors criteria. We started by using machine learning methods without any kind of sampling procedure. Lower rates of accuracy and recall are obtained because to the imbalanced nature of the dataset. We then successfully implemented the aforementioned sampling strategies.

4 EXPERIMENTATION AND RESULTS:

4.1 Results with Random over sampling:

Support Vector Machines have 99% accuracy with a recall rate of 99.7% when using a random oversampling strategy. Similarly effective algorithms include the Extra Tree Classifier and the Gradient Boosting technique. All sampling methods may benefit from this method's increased precision. However, this approach does not rely on any heuristics.

TABLE 1: RESULTS WITH RANDOM OVER SAMPLING

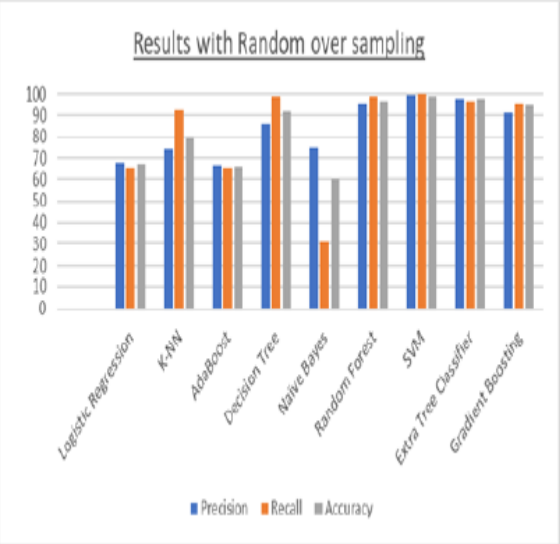

Figure 2: *Results of various algorithms with random over sampling*

TABLE 1: RESULTS WITH RANDOM OVER SAMPLING

Algorithm	Precision	Recall	Accuracy
Logistic Regression	68	66	67.5
K-NN	74	92	79.4
AdaBoost	67	66	66.6
Decision Tree	86	99	91.5
Naïve Bayes	75	31	60
Random Forest	95	99	97
SVM	99.7	100	99
Extra Tree Classifier	98	97	97.8
Gradient Boosting	91	95	94.6

4.2 Results with Synthetic Minority Oversampling:

Accuracy of 91% and recall rate of 93% are achieved using Synthetic Minority Oversampling, Random Forest, and Extra tree Classifier. Similarly effective is the Gradient Boosting algorithm. This approach employs a heuristic technique for sampling, thus although it is less accurate than random oversampling, it is still viable for real-time data samples.

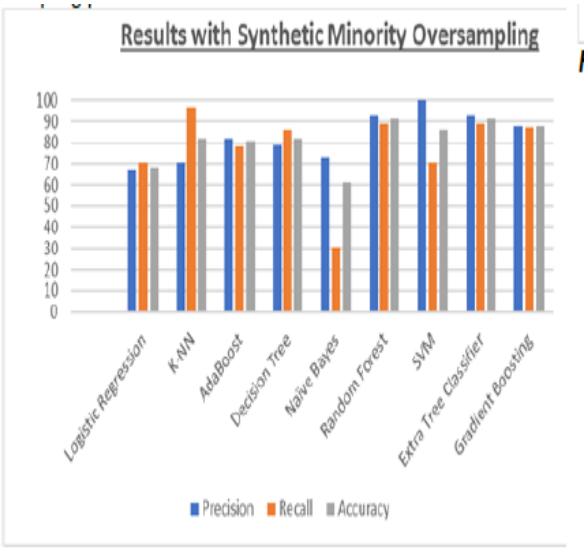

Figure 3: Results of various algorithms with SMOTE

TABLE 2: RESULTS WITH SYNTHETIC MINORITY OVERSAMPLING

Algorithm	Precision	Recall	Accuracy
Logistic Regression	67	71	68.8
K-NN	71	97	82
AdaBoost	82	78	80.8
Decision Tree	79	86	82
Naïve Bayes	73	31	61
Random Forest	93	89	91.3
SVM	100	71	86
Extra Tree Classifier	93	89	91
Gradient Boosting	88	87	87.8

4.3 Results with Adaptive synthetic sampling approach:

The combination of the Random Forest and Extra tree Classifier methods with the Adaptive synthetic sampling method yields a 90.3% accuracy and a 93% recall rate. Similarly effective is the Gradient Boosting algorithm.

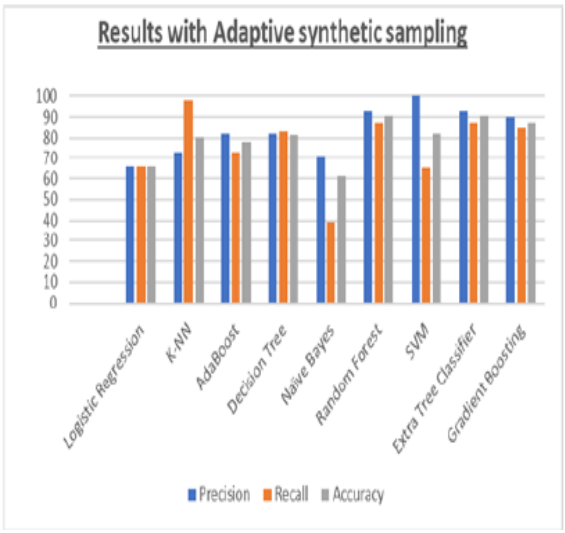

Figure 4: Results of various algorithms with ADASYN

TABLE 3: RESULTS WITH ADAPTIVE SYNTHETIC SAMPLING

Algorithm	Precision	Recall	Accuracy
Logistic Regression	66	66	65.7
K-NN	73	98	80.5
AdaBoost	82	73	78
Decision Tree	82	83	81.8
Naïve Bayes	71	39	61
Random Forest	93	87	90.3
SVM	100	65	82.3
Extra Tree Classifier	93	87	90.3
Gradient Boosting	90	85	87.4

5. CONCLUSION

In this research, we used machine learning techniques to identify cardiac problems. Since raw datasets tend to have skewed class distribution samples, we used three different sampling methods to even things out. Accuracy and recall improved dramatically once sampling methods were used. SVM had the highest precision when randomly oversampling data. The highest accuracy was achieved by Random Forest and Extra tree Classifier for Synthetic Minority Oversampling. The most accurate adaptive synthetic

sampling methods are the Random Forest and Extra tree Classifier.

REFERENCES

[1] J Thomas MR, Lip GY. Novel risk markers and risk assessments for cardiovascular disease. Circulation research. 2017; 120(1):133–149. https://doi.org/10.1161/CIRCRESAHA.116.309955 PMID: 28057790

*[2] Ahmed M. AlaaID1, Thomas Bolton, Emanuele Di Angelantonio, James H.F. RuddID, Mihaela van der Schaar,―Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants‖, PLOS ONE 14(5): e0213653. https://doi.org/10.1371/journal, May 15, 2019H. Poor, ―A Hypertext History of Multiuser Dimensions,‖ MUD History, http://www.ccs.neu.edu/home/pb/mud-history.html. 1986. (URL link *include year)*

[3] Stephen F. Weng, Jenna Reps, Joe Kai1, Jonathan M. Garibaldi, Nadeem Qureshi, ―Can machine-learning improve cardiovascular risk prediction using routine clinical data?‖, PLOS ONE | https://doi.org/10.1371/journal.pone. 0174944 April 4, 2017

[4] Rine Nakanishi, Damini Dey, Frederic Commandeur, Piotr Slomka, ―Machine Learning in Predicting Coronary Heart Disease and Cardiovascular Disease Events: Results from The Multi-Ethnic Study of Atherosclerosis (Mesa)‖, JACC Mar-20, 2018, Volume 71, Issue 11

[5] https://www.cdc.gov/heartdisease/facts.htm. Available [Online].

[6] Senthilkumar Mohan, Chandrasegar Thirumalai, Gautam Srivastava ―Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques‖, Digital Object Identifier 10.1109/ACCESS.2019.2923707, IEEE Access, VOLUME 7, 2019 S.P. Bingulac, ―On the Compatibility of Adaptive Controllers,‖ Proc. Fourth Ann. Allerton Conf. Circuits and Systems Theory, pp. 8-16, 1994. (Conference proceedings)

[7] A. Gavhane, G. Kokkula, I. Pandya, and K. Devadkar, __Prediction of heart disease using machine learning,'' in Proc. 2nd Int. Conf. Electron., Commun. Aerosp. Technol. (ICECA), Mar. 2018, pp. 1275–1278.

[8] M. Sultana, A. Haider, and M. S. Uddin, ―Analysis of data mining techniques for heart disease prediction,‖ 2016 3rd Int. Conf. Electr. Eng. Inf. Commun. Technol. ICEEICT 2016, 2017

[9] M. Akhil, B. L. Deekshatulu, and P. Chandra, ―Classification of Heart Disease Using K-Nearest Neighbor and Genetic Algorithm,‖ Procedia Technol., vol. 10, pp. 85–94, 2013.

[10] N. Al-milli, __Backpropogation neural network for prediction of heart disease,'' J. Theor. Appl.Inf. Technol., vol. 56, no. 1, pp. 131–135, 2013.