Machine Learning Techniques for Heart Disease Prediction

A. Lakshmanarao,Y.Swathi, P.Sri Sai Sundareswar

Abstract—According to WHO (World Health Organization), Heart diseases are the reason for 12 million deaths every year. In most of the countries, half of the deaths are due to cardiovascular diseases. The early diagnosis of cardiovascular sicknesses can help in settling on choices on the way of life changes in high hazard patients and thusly diminish the difficulties. In this paper, machine learning techniques are used for the detection of heart disease. We also applied sampling techniques for handling unbalanced datasets. Various machine learning methods are used to predict the overall risk. The framingham_heart_disease dataset is publically available on the Kaggle. This dataset is used in our experiments. The end goal is to predict whether the patient has a 10-year risk of future coronary heart disease (CHD). The dataset contains 15 features that give patient information. By applying machine learning techniques, we achieved 99% accuracy in heart disease detection.

Index Terms— Heart disease, CHD, Machine learning, sampling.

1 Introduction

The most common type of heart disease is Coronary heart disease (CHD). More than 3,50,000 people are dying with this disease annually. In the United States every year 610000 people are dying with heart disease. Every year 7,35,000 members having a heart attack, out of which 5,25,000 members are 1st heart attack and 2,10,000 members who have already had a heart attack [5]. In Asian countries also,22% of the deaths (in total heart disease deaths) are due to heart diseases. The contributing factors for heart disease are more (blood pressure, diabetes, current smoker, high cholesterol, etc..). So it is not easy to identify heart disease. Various techniques in data mining and neural networks have been employed to find out the severity of heart disease among humans. The nature of CHD illness is perplexing, what's more, subsequently, the malady must be taken care of cautiously. Not doing early detection, may influence the heart or cause unexpected passing. The point of view of medicinal science and information digging are utilized for finding different sorts of metabolic disorders. Machine Learning is a technique that helps the system to learn from previous data samples, examples without being explicitly programmed. Machine learning creates logic based on historical data. Machine Learning plays a vital role in many fields. It also shows its impact on heart disease detection. Deep Learning is a part AI, which can also be considered as a subset of machine learning. Deep learning can also be applied to the number of research areas. It is also applied for heart disease prediction.

- *A. Lakshmanarao, Assistant Professor, Dept of CSE, Raghu Engineering College, Dakamarri, Visakhapatnam, Email: laxman1216@gmail.com*
- *Y. Swathi, Assistant Professor Dept of CSE, BABA Institute of Technology & Sciences, Visakhapatnam, Email: swathiyendamuri@gmail.com*
- *P. Sri Sai Sundareswar, B. Tech Student, Dept of CSE, Raghu Engineering College, Dakamarri, Visakhapatnam, Email: psundareswar@gmail.com*

2 Literature Survey

The leading cause for mortality and morbidity is cardiovascular disease [1]. Ahmed M. Alaa[2] et.al proposed machine learning techniques for Cardiovascular disease risk prediction. But they achieved maximum accuracy of 77%. As the dataset is unbalanced, there is a need to apply sampling techniques. But they directly applied Machine learning models on the dataset.Stephen F. Weng[3] et.al studied application of machine learning algorithms to improve cardiovascular risk prediction. They shown that Machine-learning algorithms are successful in improving accuracy of cardiovascular risk prediction, but the required number of patient records must be more to achieve better results. Rine Nakanishi [4] et.al evaluated ML methods for improving the prediction rate of coronary heart disease (CHD). They applied machine learning approaches on 6814 patient records and achieved good accuracy rate. Senthilkumar Mohan[6] proposed a machine learning model that finds significant features for improving the prediction rate of cardiovascular disease. They tried with various combinations of features and achieved an accuracy of 88.7% with hybrid random forest. Himanshu Sharma [7] et.al applied K-Nearest Neighbor Algorithm (KNN), Decision Trees (DT), SVM, Naive Bayes algorithms for heart disease prediction and achieved good results.

Marjia Sultana [8] et.al have explored the role of datasets accessibility for Heart disease illness are commonly raw in nature which is profoundly repetitive and conflicting. There is a need for pre-handling of these datasets before applying machine learning techniques. They also suggested that the selection of crucial features plays a vital role in achieving a good accuracy rate. M.A.Jabbar [9] et.al proposed a method which specifies the importance of selection of features in heart disease prediction. They applied Genetic algorithm for feature selection and later applied K-NN and achieved good results.Some of the researchers also applied deep learning techniques for heart disease prediction. N. Al-milli [10] proposed a deep learning method with 13 features. Their results show an enhanced level of accuracy when compared with other techniques.

3 PROPOSED METHOD

We collected a dataset from Kaggle for heart disease prediction. The dataset contains 4239 patients' records with 15 features. Features include medical risk factors, demographic, behavioral factors. Features are age, sex, education, current smoker, cigsPerDay, BPMeds, prevalentStroke, prevalentHyp, diabetes, totChol, sysBP, diaBP, BMI, heartrate, glucose. Based on these features, we need to predict whether the patient has a 10-year risk of future coronary heart disease (CHD). Dataset contains 644 samples of TenYearCHD as 1 and remaining samples with TenYearCHD as 0.

Data preprocessing is an important step in machine learning. When dealing with unbalanced datasets, oversampling and undersampling are helpful to balance the samples of two different classes. As the dataset is unbalance, we applied three sampling techniques on the dataset.

i) Random Over sampling
ii) Synthetic Minority Oversampling
iii) Adaptive synthetic sampling approach

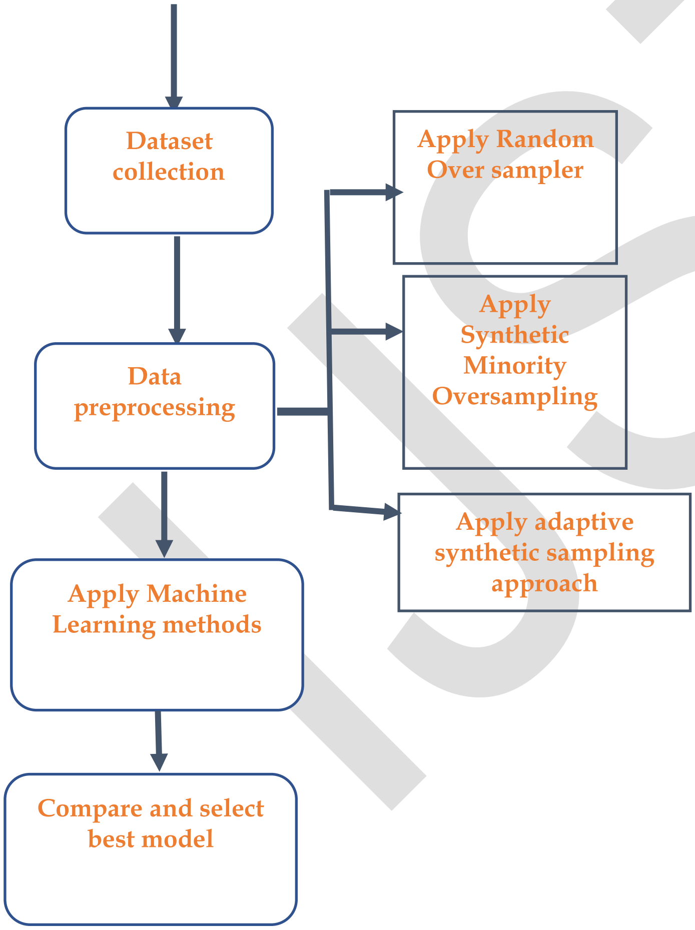
Dataset collection
Data preprocessing
Apply Machine Learning methods
Compare and select best model
Apply Random Over sampler
Apply Synthetic Minority Oversampling
Apply adaptive synthetic sampling approach

Figure 1: Proposed model

3.1 Random over sampling:

This oversampling technique is utilized to supplement the preparation information with different duplicates of a portion of the minority classes. This oversampling should be possible more than once. This is one of the basic techniques, that is likewise demonstrated to be vigorous. Instead of copying each example in the minority class, some of them might be arbitrarily picked with substitution. In simple terms, existing samples are sampled with replacement for producing new samples.

3.2 Synthetic Minority Oversampling (SMOTE):

SMOTE uses specific heuristic method instead of repeating same samples. Consider some training data which has n samples, and f features in the feature space of the data. Note that these features, Consider n training data samples (with f-number features).In this technique, it takes a sample from the given dataset and k is the number of nearest neighbors. It considers a vector between present information point and one of the k neighbors for creating synthetic data point and increased the vector by number x which is in the range of 0 and 1.This pint is added to create a new point.

3.3 Adaptive synthetic sampling approach (ADASYN):

ADASYN also uses heuristic method. It is based on SMOTE. It shifts the classification boundary to minority which are difficult. It uses weighted distribution for minority class samples based on their level of difficulty in learning. The difference between SMOTE and ADASYN is that, ADASYN mainly concentrate on the data samples which are difficult to classify with a nearest-neighbors rule where as SMOTE will not make any distinction.

First, we applied machine learning techniques without sampling techniques. As the dataset in unbalanced, we get less accuracy rate and recall rate. After that we applied above sampling techniques and achieved good results.

4 EXPERIMENTATION AND RESULTS:

4.1 Results with Random over sampling:

With random oversampling technique, Support Vector machine gives an accuracy of 99% and recall rate of 99.7%. ExtraTree Classifier and Gradient Boosting algorithms also gives good results.This technique gives more accuracy in all sampling techiques.But,this technique is not based on any heuristic method.

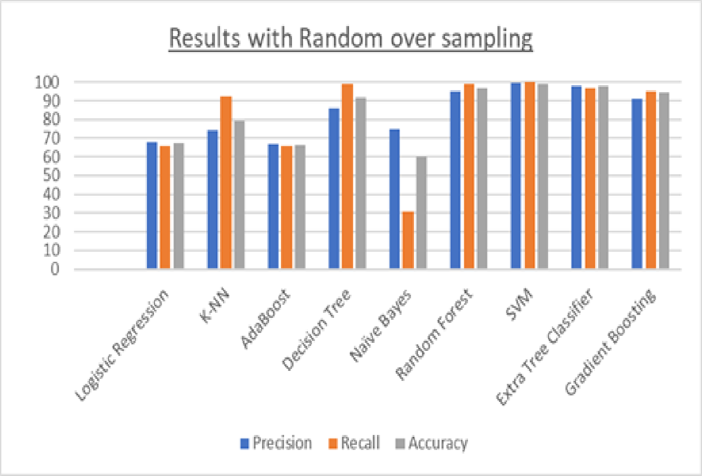

Figure 2: Results of various algorithms with random over sampling

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:

With Synthetic Minority Oversampling technique, Random Forest and Extratree Classifier techniques gives an accuracy of 91% and recall rate of 93%. Gradient Boosting algorithm also gives good results. Although this method gives less accuracy than random oversampling, it is considered for real-time data samples because it uses heuristic method for sampling process.

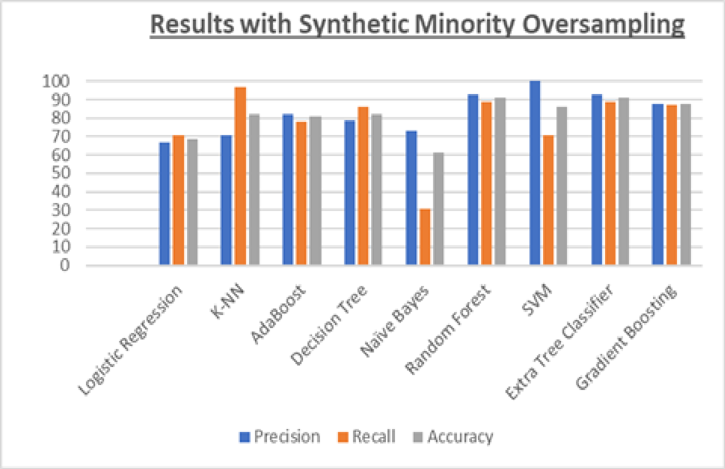

Figure 3: Results of various algorithms with SMOTE

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:

With Adaptive synthetic sampling technique, Random Forest and Extratree Classifier techniques gives an accuracy of 90.3% and recall rate of 93%. Gradient Boosting algorithm also gives good results.

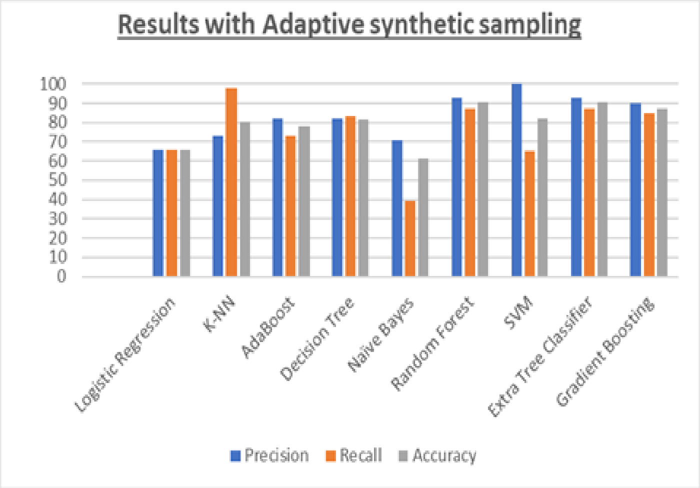

Figure 4: Results of various algorithms with ADASYN

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 paper, we applied machine learning methods for heart disease detection. As raw datasets contain unbalanced samples of class distribution, we applied three sampling techniques on the dataset. After applying sampling techniques accuracy and recall rates increased drastically. For random oversampling, SVM given the best accuracy. For Synthetic Minority Oversampling, Random Forest and Extratree Classifier given the best accuracy. For Adaptive synthetic sampling, Random Forest and Extratree Classifier given the best accuracy.

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.