HEART DISEASE PREDICTION USING MACHINE LEARNING TECHNIQUES

Ravula Bala Ranga sai[1], Valluri Satish[2], Chennuboina Purna Sekhar[3], Sathiri Karthik[4]

[1] *UG Student, Dept. of Electronics and Communication Engineering, Vasireddy Venkatadri Institute of Technology, Nambur, Andhra Pradesh, India*

[2] *UG Student, Dept. of Electronics and Communication Engineering, Vasireddy Venkatadri Institute of Technology, Nambur, Andhra Pradesh, India*

[3] *UG Student, Dept. of Electronics and Communication Engineering, Vasireddy Venkatadri Institute of Technology, Nambur, Andhra Pradesh, India*

[4] *UG Student, Dept. of Electronics and Communication Engineering, Vasireddy Venkatadri Institute of Technology, Nambur, Andhra Pradesh, India*

ABSTRACT

Heart Disease is currently the top cause of mortality worldwide. Several strategies have been developed by researchers to increase the sharpness and efficiency of clinical cardiac disease detection. Moreover, for the people who study clinical data, predicting of cardiac disease remains a challenging task since it must be predicted in its early stage for a person to survive. For its prediction in early stage the data in the dataset must be balanced.

The major goal of this project is to develop an accurate Cardiac Disease prediction model that uses the NEARMISS under sampling algorithm for balancing the imbalanced distribution of the data present in dataset. In machine learning, the term "data imbalance" refers to an unbalanced distribution of classes within a dataset. This problem mostly arises in classification jobs. The reason behind it is distribution of labels within a dataset is not symmetrical. So, in order to overcome it near miss under sampling technique is used for improving the performance of the models (Random Forest (RF), Logistic Regression (LR), Support Vector Machine(SVM), Decision Tree(DT), K-Nearest Neighbors(KNN) and XG-Boost) in terms of accuracy and recall.

Keywords: *Logistic Regression (LR), Decision Tree (DT), Kth Nearest neighbors (KNN), Support Vector Machine (SVM), Random Forest (RF), XG-Boost, Near Miss, Data Imbalance, Machine Learning.*

---

1. INTRODUCTION

The top cause of death worldwide, as reported by the WHO, is heart disease [1]. It is estimated that more than 17.9 million people die each year from cardiovascular disorders, with coronary artery disease and cerebral stroke being responsible for 80% of these deaths. Heart disease is frequently determined by several risk factors, including smoking, excessive alcohol and caffeine use, stress, and physical inactivity, as well as physiological variables including obesity, hypertension, high blood cholesterol, and pre-existing cardiac diseases. Taking preventative actions to avoid the consequences that such diseases cause is highly dependent on the effective, accurate, and early medical examination of heart disease. Dealing with issues with complex and nonlinear characteristics, machine

learning algorithms have a clear edge. Many algorithms, including LR, SVM, and KNN, among others, have been effectively used to solve several illnesses classification and prediction difficulties, including early warning for Electrocardiogram (ECG) detection [2] and prediction linked to congenital heart disease [3].

To determine the best accurate model, several machine learning algorithms are evaluated in terms of accuracy and recall scores, including Logistic Regression, Support Vector Machine, K Nearest Neighbor, Decision Tree, Random Forest, and the ensemble approach of XG-Boost. Here, the website Kaggle's heart disease dataset is used.

2. LITERATURE SURVEY

- **Gupta et al. (2020):** [4] Created a machine intelligence framework consisting of factor analysis of mixed data (FAMD) and RF-based MLA. The necessary tr traits were discovered using the FAMD, and the disease was predicted using the RF. The results of the experiments demonstrated that the suggested approach performed better than other models and the findings of earlier studies, obtaining accuracy, sensitivity, and specificity of up to 93.44%, 89.28%, and 96.96% respectively.

- **Archana Singh. (2020):** [5] Using data from the UCI repository for training and testing, they estimated the accuracy of machine learning methods for the prediction of heart disease. These algorithms included k-nearest neighbor, decision trees, linear regression, and support vector machines (SVM). They achieved 87% accuracy with KNN and 83% accuracy with SVM.

- **Bugra Kaan Türkmenoğlu. (2021):** [6] Because of the data set's uneven class distribution, Resampling techniques were used. Experimental investigations have demonstrated that applying data cleaning and resampling combined increases the success of predictions. It was shown that removing the class imbalance from the data set improved the classifier's performance. While the under-sampling technique had a higher success rate with the Extra Trees algorithm (%84.58), the oversampling method had a better success rate with the Random Forest algorithm (%84.51).

- **Ch Raja Shaker.** (**2022):** [9] In the diagnosis of cardiovascular illnesses, these machine learning algorithms outperformed deep learning. The 11 fields in the dataset have each had their relative importance evaluated using the PCA approach. Accuracy and recall rates rose when sample techniques were used. The data show that Naive Bayes, Decision Tree, and Random Forest classifiers outperform other ML techniques.

- **Arsalan Khan. (2023): [10]** For the categorization and forecasting of CVD patients in Pakistan, ML techniques such decision trees (DT), random forests (RF), logistic regression (LR), Naive Bayes (NB), and support vector machines (SVM) were used. For each method, they ran exploratory analysis and experimental output analysis. For each algorithm, they also estimated the confusion matrix and recursive operating characteristic curve. The performance of the recommended machine learning algorithm was estimated under various circumstances in order to choose the best suited machine learning algorithm within the class of models. For CVD, the RF algorithm's prediction accuracy, sensitivity, and recursive operational characteristic curve were the highest at 85.01%, 92.11%, and 87.73%, respectively. Moreover, it had misclassification errors for CVD of 43.48% and 8.70%, respectively, with the lowest specificity. Our findings demonstrated that the RF algorithm is the effective method for classifying and predicting CVD.

3. PROBLEM IDENTIFICATION

Finding a cure for cardiac disease is a huge task. Although there are tools that can forecast heart disease, they are either prohibitively expensive or ineffective when it comes to calculating the likelihood of cardiac disease in humans. The death rate and total consequences can be reduced by heart illnesses that are detected early. In the modern world, we have a lot of data, so we can use a variety of machine learning algorithms to examine the data and look for hidden patterns. With medical data, the hidden patterns might be exploited for health diagnosis. In the

medical industry, the use of data mining can lead to the identification and extraction of interesting patterns and information that can be used for making clinical diagnoses. Using an imbalanced data set is a further issue that is associated to this. In machine learning, the term "data imbalance" refers to an uneven distribution of classes within a dataset [11]. This problem mostly arises in classification jobs if the distribution of classes or labels within a dataset is not symmetrical. The resampling approach, which involves adding records to the minority class or removing ones from the majority class, is the simple solution to this Problem.

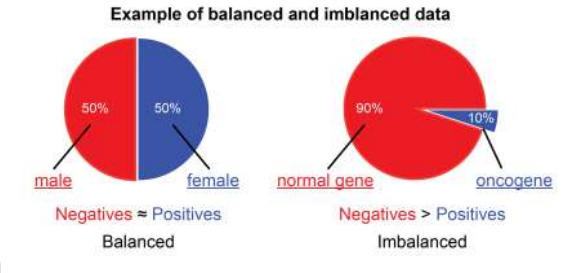

Fig.1. Balanced Vs Imbalance Data

In machine learning, data imbalance can lead to several issues, which includes: Unbalance: The unbalanced models may overestimate the majority class and underestimate the minority class. This might be an issue in circumstances when the cost of false negatives is considerable, such in medical diagnostics. Unsatisfactory generalization: Imbalanced data might cause models to perform well on the training set but badly on the test set since they are unable to adapt adequately to new data. Overfitting: Imbalanced data can result in overfitting, in which the model matches the training data too closely and is unable to generalize to new data. Absence of model assessment: With unbalanced datasets, traditional evaluation metrics like accuracy might be deceptive. A model that consistently forecasts the majority class, for instance, could have great accuracy, but it is useless for forecasting the minority class.

4. PROPOSED METHODOLOGY

4.1 Data Set:

Table – 1: Dataset

Attributes	Description	Range
Age	Age of person in years	29-79
Gender	Gender of person (1-M, 0-F)	0,1
Cp	Chest pain type	1,2,3,4
Testbps	Resting blood pressure in mm Hg	94-200
Chol	Serum cholesterol in mg/dl	126-564
Fbs	Fasting blood sugar in mg/dl	0,1
Restecg	Resting ECG results	0,1,2
Thalach	Maximum heart rate achieved	71-202
Exang	Exercise Induced Angina	0,1
OldPeak	ST depression induced by exercise relative to rest	1-3
Slope	Slope of the peak Exercise ST segment	1,2,3
Ca	Number of vessels colored by fluoroscopy	0-3
Thal	3 – Normal, 6 – Fixed Defect, 7 – Reversible Defect	3,6,7
Result	Class Attribute	0,1

This dataset includes patients ranging in age from 29 to 79. Patients who are male are represented by the gender value 1 and those who are female by the gender value 0. There are four forms of chest discomfort that are thought to be signs of heart disease. Because of clogged coronary arteries, type 1 angina is brought on by decreased blood supply to the heart muscles. While under mental or emotional stress, type 1 angina, a chest discomfort, develops. Chest discomfort that is not caused by angina might have many different causes and isn't always the result of a heart condition. The fourth category, Asymptomatic, could not be a heart disease sign. The resting blood pressure measurement is the following characteristic, trestbps. The cholesterol level is chol. Fbs stands for fasting blood sugar level; a value of 1 is given if it is less than 120 mg/dl and a value of 0 if it is more. Exang is the exercise-induced angina, which is recorded as 1 if there is pain and 0 if there is none, and Restecg is the resting electrocardiographic result. Oldpeak is the exercise-induced ST depression, slope is the exercise-induced ST segment peak slope, ca is the number of main vessels stained by fluoroscopy, thal is the exercise test time in minutes, and num is the class attribute. Patients with heart disease have a value of 1 for the class attribute, while normal people have a value 0.

4.2 Under Sampling:

Under sampling [12] is a method of balancing the class distribution in a classification dataset having a skewed class distribution. Under sampling eliminates cases from the training dataset that belong to the dominant class in order to balance the class distribution, lowering the skew from a 3:300 to a 3:30, 1:2, or even a 1:1 class distribution. This article employed an under-sampling technique based on the Near Miss method to assess the impact of the data-point method. Given its benefits in providing a more reliable and equitable class distribution boundary, Near Miss was chosen since it was proven to enhance the performance of classifiers for detection in sizable unbalanced Datasets.

4.2.1 Near Miss Algorithm

The Near Miss algorithm selects samples from the positive class that are close to those in the negative class and is a popular under sampling technique in machine learning.

1. Split the dataset into minority and majority classes: X_minority = {x_i:y_i = 1},X_majority = {x_i:y_i = 0}
2. Calculate the number of examples in the majority class: n_majority = len(X_minority)
3. Calculate the number of examples in the majority class: n_majority = len(X_majority)
4. Calculate the ratio of the minority to the majority class: ratio = n_minority / n_majority
5. Initialize an empty set to story the selected examples: X_selected = {}
6. For each example x_i in X_minority:
    a. Find the k nearest neighbors of x_i in X_majority.
    b. Add the K nearest neighbors to X_selected
7. Return X_selected

5. RESULTS

In this work, Near Miss-based approach for processing the imbalanced datasets for heart disease prediction, this method is applied for several machine learning classification algorithm models to increase accuracy and recall scores. The results obtained from this work demonstrated that the proposed model works well for Recall scores. Among which the Random Forest algorithm surpassed all others, with an accuracy of 90.16% and a recall value of 93.75%. Nevertheless, KNN's performance is the lowest of all, with an accuracy and recall value of 67.21% and 73.33%, respectively.

Table – 2: Accuracy & Recall Values Before Applying Near-Miss

Model	Accuracy	Recall
KNN	67.97%	71.88%
XG-Boost	78.69%	78.69%
SVM	81.97%	81.08%
Decision Tree	81.97%	84.85%
Logistic regression	85.25%	85.71%
Random forest	90.16%	88.69%

Table – 3: Accuracy & Recall Values After Appling Near-Miss

Model	Accuracy	Recall
KNN	67.21%	73.33%
XG-Boost	78.69%	84.85%
SVM	81.97%	88.24%
Decision Tree	81.97%	79.41%
Logistic regression	85.25%	85.71%
Random forest	90.16%	93.75%

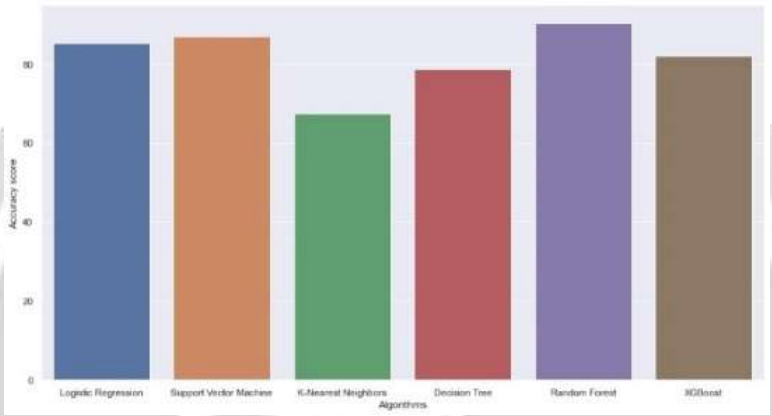

Fig - 2: Comparison of Accuracy

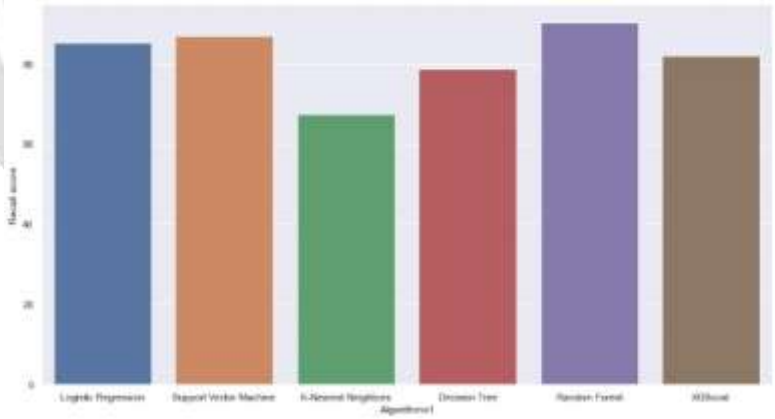

Fig – 3: Comparison of Recall

6. CONCLUSION

By combining the near miss under sampling technique with Random Forest based machine learning algorithm, an improved heart disease prediction model with higher accuracy and recall values were developed. The aim of this project is to develop an efficient and effective CVD prediction methods using machine learning. For this the data present in dataset should be balanced so the unbalanced training dataset is balanced using the Near Miss technique, and various machine learning algorithms were employed to construct the prediction model. Before

applying Near-Miss under sampling technique to the Dataset, Random Forest achieved accuracy of 90.16% and recall of 88.89% thereby after applying the Nera miss algorithm Random Forest gave its outstanding results of accuracy 90.16% and recall of 93.75% among the other machine learning techniques.

In Future the performance of the model's can be further improved by expanding the quantity of the dataset that is by integrating genetical data and wearable devices data, feature selection, and Hyperparameter tuning, as well as hybrid sampling strategies which may result in real-time and more accurate cardiac disease detection prototypes.

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] Che, C.; Zhang, P.; Zhu, M.; Qu, Y.; Jin, B. Constrained transformer network for ECG signal processing and arrhythmia classification. BMC Med Informatics Decis. Mak. 2021, 21, 184.

[3] Hoodbhoy, Z.; Jiwani, U.; Sattar, S.; Salam, R.; Hasan, B.; Das, J. Diagnostic Accuracy of Machine Learning Models to Identify Congenital Heart Disease: A Meta-Analysis. Front. Arif. Intel. 2021, 4, 197.

[4] A. Gupta, R. Kumar, H. S. Arora, intelligence framework for heart disease diagnosis, ''IEEE Access, vol. 8, pp. 14659–14674,2020.

[5] Angraal S, Mortazavi BJ, Gupta A, Khera R, Ahmad T, Desai NR, Jacoby DL, Masoudi FA, Spertus JA, Krumholz HM, "Machine Learning Prediction of Mortality and Hospitalization in Heart Failure With Preserved Ejection Fraction", JACC : Heart Failure, vol. 8, Issue 1, January 2020

[6] B. K. Turkmenoglu and O. Yildiz, "Predicting the survival of heart failure patients in unbalanced data sets," in *2021 29th Signal Processing and Communications Applications Conference (SIU)*, pp. 1–4, Istanbul, Turkey, 2021. View at: Google Scholar

[7] A. Ishaq, S. Sadiq, M. Umer et al., "Improving the prediction of heart failure patients' survival using SMOTE and effective data mining techniques," *IEEE Access*, vol. 9, pp. 39707–39716, 2021. View at: Publisher Site | Google Scholar

[8] Ch Raja Shaker, Anisetti Sidhartha, Anto Praveen, A.chrsity, B.Bharati, "An Analysis of Heart Disease Prediction using Machine Learning and Deep Leaning Techniques", 2022 6th International Conference on Trends in Electronics and Informatics(ICOEI),pp.1484-1491, 2022.
a. View at: Show Article | Google Scholar

[9] Arsalan Khan, Moiz Qureshi, Muhammad Daniyal, Kassim Tawiah, "A Novel Study on Machine Learning Algorithm Based CVD Prediction", Health & Social Care in the Community, Vol.2023.
a. View at: CrossRef | Google Scholar

[10] Roweida Mohammed, Jumanah Rawashdeh, Malak Abdullah, "Machine Leaning with Oversampling and Under Sampling Techniques. DOI: 10.1109/ICICS49469.2020.239556

[11] Nh.Lakanipho Michael Mquadi, Timothy Adeliyi solving, "Misclassification Of The Credit Card Imbalance Problem Using Near Miss". Volume 2021 | Article ID 7194728 | https://doi.org/10.1155/2021/7194728.