2024 Global Conference on Communications and Information Technologies (GCCIT)
Bangalore, India. Oct 25-26, 2024

Effect of Different Oversampling Techniques to Handle Class Imbalance Challenges in Coronary Heart Disease Prediction

Amiya Ranjan Panda
School of Computer Engineering
KIIT Deemed to be University
Bhubaneswar, India
amiya.pandafcs@kiit.ac.in

Shatadru Banerjee*
School of Computer Engineering
KIIT Deemed to be University
Bhubaneswar, India
2105580@kiit.ac.in

Shashwat Naik
School of Computer Engineering
KIIT Deemed to be University
Bhubaneswar, India
2105490@kiit.ac.in

2024 Global Conference on Communications and Information Technologies (GCCIT) | 979-8-3503-8891-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/GCCIT63234.2024.10862873

***Abstract*—CHD or Coronary Heart Disease has a considerable influence on global mortality rate, hence ahead of time and precise prediction is crucial for prompt treatment and avoid patient deaths. However, predicting CHD is difficult since the dataset is skewed. To address this challenge, several oversampling techniques are proposed, including Synthetic Minority Oversampling Technique (SMOTE), random oversampling, and Adaptive Synthetic (ADASYN). Incorporating these strategies lowers classifier predisposition towards the majority group while emphasizing proper knowledge of minority notions. Most algorithms perform better on balanced data rather than imbalanced data.**

This article tests six machine learning models using the CDG dataset. Because this is an imbalanced data problem, accuracy is not used as the primary metric. F1-score is the primary metric. The Extra Trees Classifier achieves an F1-score of 95% by employing random oversampling techniques. F1-score, accuracy, precision, Matthews Correlation Coefficient(MCC) and recall are used for measuring the performance of each algorithm.

Keywords— CHD, class imbalance, oversampling, heart disease prediction, machine learning

I. Introduction

Coronary heart disease (CHD) significantly contributes to impairment and deaths associated with cardiovascular diseases (CVDs), and it stands as the principle mortality factor globally. CVDs killed 18 million people in 2019, representing 32% (one-third) of all fatalities globally. Strokes and heart attacks accounted for almost 85% of these fatalities [1]. CHD can cause thoracic distress or unease (angina), breathlessness, and even a cardiac arrest when the arteries that supply oxygen and blood to the heart narrow or become clogged. Age, gender, smoking, alcohol consumption, sleep duration, asthma, kidney disease, skin cancer, diabetes, and BMI are just a few of the many variables that may influence this complex disease. Locating patients with a high likelihood of developing CHD is a critical task in predicting CHD. It is critical for early disease detection, and cure [2]. Identifying CVD can be challenging even for medical experts. Using an automated system can improve CVD prediction accuracy [3][4].

The utilization of machine learning (ML) within the healthcare sector has lately increased [5]. Machine learning algorithms have been used in a variety of real-world circumstances, such as cardiac arrest prediction, COVID prediction, lung cancer classification, and so on, to aid us make precise predictions [6][7]. By analyzing clinical and population data, machine learning algorithms have generated promising results in forecasting the risk of heart disease. Logistic Regression, Random Forests, Boosting algorithms and Decision Trees are the most common algorithms for machine learning used in CHD prediction [8]. However, an imbalance in class distribution makes it significantly difficult to predict CHD, owing to the lower number of positive cases (CHD patients) compared to negative cases (patients without CHD).

This uneven distribution influences machine learning algorithm performance, potentially leading to a biased model that incorrectly classifies patients with Coronary Heart Disease (CHD) as non-CHD, causing delays in diagnosis and treatment. The objective of our experiment is to pinpoint the influence of oversampling to handle imbalance data to enhance the performance of machine learning algorithms for the prediction of cardiovascular disease (CVD). Section II presents some related studies in this area. Section III describes the dataset and models . Section IV presents the outcomes of the experiment, while Section V represents the conclusion and future works on this topic.

II. Literature Survey

Several studies have been carried out to predict coronary heart disease. Various data sets are utilized in the analysis. Traditional machine learning algorithms are commonly used in conjunction with ensemble methods to improve classification accuracy. However, almost every study used small datasets, which are insufficient to predict things like heart disease.

In 2021, Rohit Bharti et al. proposed predicting coronary heart disease through the integration of machine learning and deep learning [9]. The investigation utilized a dataset for predicting heart disease from the UCI repository. The Lasso feature selection technique was used. SVM, LR, RF, DT, KNN, and Deep Learning all achieved 83%, 85%, 80%, and 94% accuracy, respectively.

Lakshmi et al. [10] suggested a method for identifying the optimal collection of diagnostic criteria that combines classic Machine learning algorithms with state-of-the-art gradient boosting methods. In this suggested system, genetic algorithm based feature selection strategy reduces the quantity of variables by 20% while retaining model accuracy.

In 2020, Pranov Motarwar et al. proposed an intellectual methodology for predicting heart disease through machine learning [11]. Their study was based on the Cleveland dataset.

The suggested study used five algorithms: GNB, RF, SVM, Logistic Model Tree, and Hoeffding DT. They used different techniques for feature selection. The experiment estimated accuracy of 95% for RF, 90% for SVM, 93% for GNB, 81% for LMT, and 81% for Hoeffding DT.

Santhana Krishnan. J et al. implemented DT and NB machine learning algorithms in their study [12]. They achieved 91% accuracy for DT and 87% for NB.

Yilmaz and Yagm [13] compared the effectiveness of machine learning techniques for predictive categorization of CHD. Three distinct methods were formulated using RF, LR, and SVM techniques.

Shaik Farzana et al used various machine learning classification techniques, including RF, KNN, GNB, SVM, and XGBoost [14]. They used the UCI dataset and estimated accuracy for RF (89%), KNN (67%), GNB (82%), SVM (82%), and XGBoost (79%).

III. Proposed Methodology

Five machine learning algorithms were used to classify data in this study. The models used included Logistic Regression (LR), Random Forest Classifier (RF), Extra Trees Classifier (ET), Linear Discriminant Analysis (LDA) and Stochastic Gradient Descent (SGD).

Initially, the models were tested on imbalanced data, highlighting the challenges posed by class imbalance. To address this issue and improve the F1-score, three distinct methods of oversampling were used: random oversampling, SMOTE (Synthetic Minority Oversampling Technique), and ADASYN. These techniques helped to increase the F1-score, which was selected as the primary metric for evaluating performance of the model due to the imbalance problem. Fig. 1 shows the proposed methodology.

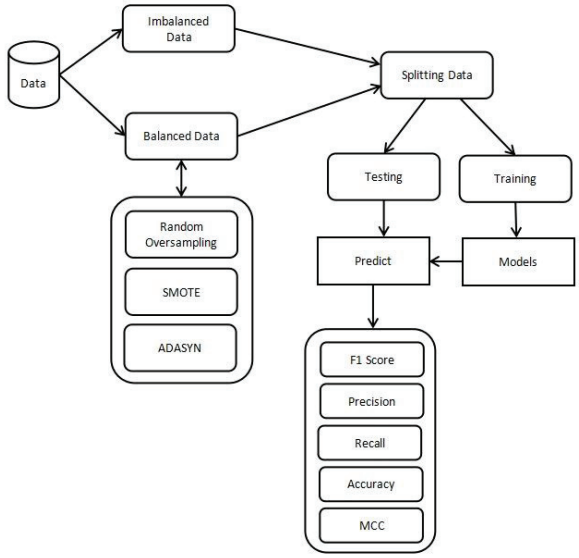

Fig. 1. Flowchart depicting the proposed methodology

After completing the steps of detecting outliers and normalization, the dataset was divided at random into training and testing sets. To determine their performance, the models were trained on the training data and then tested on the testing data. While the initial results looked promising, the algorithms' F-1 scores could be improved further.

A. Dataset Description

The Personal Key Indicators of Heart Disease dataset consists of 320K rows and 18 columns. This is a processed and condensed version of the 2020 annual CDC (Centers for Disease Control and Prevention) data from surveys conducted on 400,000 adults. The health status of each patient (row) is shown. The data was gathered through telephone surveys. The dataset contains 18 attributes, namely Body Mass Index (BMI), HeartDisease, AlcoholDrinking, Smoking (smoked at least 100 cigarettes in their lifetime), Physical/Mental Health, Stroke, Diff.Walking (difficulty walking or climbing stairs), age category, sex, diabetes, race, physical activity, sleep time, general health, asthma, skin cancer, and kidney disease.

B. Correlation matrix

The variables/attributes used in the analysis are correlated. Fig. 2 displays a matrix depicting the relationship. The matrix indicates that the correlation between the variables is very low.

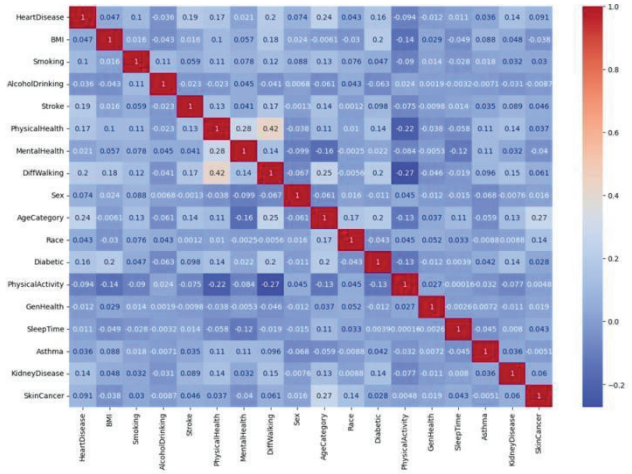

Fig. 2. Correlation matrix of the attributes in the dataset

C. Machine Learning Models

1) Random Forest Classifier

The Random Forest Classifier is an algorithm in machine learning which builds numerous decision trees using different portions of the training data [15]. In a random forest model, each tree in the forest contributes to a vote for a class, and the category that receives the most votes across all trees becomes the prediction of the model. This approach helps in improving the predictive accuracy and controlling overfitting, as it integrates the strengths of a series of decision trees.

2) Extra Trees Classifier

The Extra Trees Classifier is a type of ensemble learning method which combines the predictions from a multitude of decision trees to form a single classification outcome [16]. It shares similarities with the Random Forest Classifier but distinguishes itself in its methodology of constructing the decision trees within the "forest". In the Extra Trees method, each decision tree is created from the original training dataset. During the construction of these trees, at every split in the tree, a randomly selected portion of 'k' features is chosen from the total features available. The decision trees then determine the most suitable attribute to segment the data based on a mathematical standard, such as the Gini Index. This technique of random attribute selection contributes to the formulation of a variety of de-correlated decision trees within the forest.

3) Logistic Regression

Logistic regression algorithm is a data-driven method applied to predict the likelihood of a binary outcome using one or more predictor variables. It's specifically useful for classification tasks where the aim is to identify if a characteristic or outcome is present or absent. The model outputs probabilities that are then transformed into class predictions. Logistic regression is valued for its simplicity and interpretability, especially when explaining the connection between the independent factors and the dependent binary variable [17].

4) Stochastic Gradient Descent

Gradient descent is an enhancement method used in machine learning to minimize a cost function through a repetitive process of moving towards the minimum value. The process involves updating the parameters of the model incrementally in the opposite direction of the gradient of the cost function relative to the parameters. Stochastic Gradient Descent (SGD) is a variant of gradient descent that adjusts parameters by utilizing the gradient of the cost function with respect to a single randomly selected data point, rather than the sum of the gradients of all data points [18]. This approach can lead to faster convergence since it updates the parameters more frequently.

5) Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) serves as a statistical technique that is primarily utilized for both classification and dimensionality reduction tasks. In supervised learning scenarios, LDA seeks to discover the linear combination of attributes that optimally differentiates between two or more classes within a dataset. This method is particularly effective when the classes are well-separated and the data is approximately normally distributed.

D. Oversampling Techniques Used

1) Random Oversampling

Random oversampling is a method employed to rectify the issue of disproportionate classes in datasets, which occurs when the instances of one class significantly outnumber those of another. This imbalance can lead to predictive models that are biased towards the predominant class and perform poorly on the less represented category. By arbitrarily duplicating instances of the less represented class, random oversampling aims to balance the different categories and thus improve the model's performance on minority class predictions [19].

2) Synthetic Minority Oversampling Technique (SMOTE)

SMOTE, or the Synthetic Minority Oversampling Technique, is a prominent strategy for addressing imbalances in classes within machine learning datasets, especially for classification tasks. Differing from random oversampling, which merely replicates instances of the less represented category, SMOTE generates new, synthetic instances by estimating intermediate values between existing ones in the less represented class [20]. This technique helps to create a more evenly distributed dataset, which can lead to improved classifier performance.

3) Adaptive Synthetic Sampling (ADASYN)

ADASYN, or Adaptive Synthetic Sampling, is a technique used to tackle the challenge of class imbalance in machine learning, particularly for classification problems. It builds upon the foundation laid by SMOTE by focusing on generating synthetic data for those instances of the minority class that are harder to classify. This approach aims to improve the learning algorithm's ability to handle imbalanced datasets by providing a more diverse set of examples from the underrepresented class [21].

IV. Results

A. Evaluation Metrics

1) F-1 score

The F1-score acts as an integrated standard of measurement that merges the precision and recall of the model, calculated as their harmonic mean. This metric is especially pertinent as the chief measure in situations of class imbalance, due to its balanced evaluation of the model's precision and recall [22].

$$\mathrm{F-1\ score} = 2 \times \frac{\mathrm{Precicion} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP}\ \mathrm{FN}} \quad (1)$$

2) Precision

Precision is calculated by the proportion of accurate predictions made by the model out of all its predictions, expressed as a percentage. This metric reflects the fraction of true positive predictions in relation to the number of positive forecasts made.

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \quad (2)$$

3) Recall

Recall is measured by the fraction of true positives recognized by the algorithm, divided by the actual count of positives within the dataset. It quantifies the model's capacity to capture all relevant instances [23].

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \quad (3)$$

4) Accuracy

The accuracy rate, also known as model prediction accuracy, is quantified as the proportion of samples that the classifier has correctly identified, divided by the overall count of samples in the set [24]. It can be expressed as:

$$\mathrm{Accuracy} = \frac{(\mathrm{TP} + \mathrm{TN})}{(\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN})} \quad (4)$$

5) Matthews's Correlation Coefficient (MCC)

The Matthews Correlation Coefficient (MCC) is recognized as the most informative single-value metric for summarizing the performance of a classification model as represented in a confusion matrix. It considers true positives, true negatives, false positives, and false negatives, providing a balanced measure that is particularly useful even when the classes have significantly different sizes.

$$\mathrm{MCC} = \frac{\mathrm{TN} \times \mathrm{TP} - \mathrm{FN} \times \mathrm{FP}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \quad (5)$$

$$\mathrm{TP} = \mathrm{True\ positive}$$
$$\mathrm{TN} = \mathrm{True\ negative}$$
$$\mathrm{FP} = \mathrm{False\ positive}$$
$$\mathrm{FN} = \mathrm{False\ negative}$$

6) Confusion Matrix

A confusion matrix is a N x N matrix that evaluates the accuracy of a classification algorithm, with (N) denoting the count of distinct classes. It cross-tabulates the original class

labels with those forecasted by the model, offering a detailed breakdown of true and false predictions for each class. This matrix is instrumental in diagnosing the specific categories of classification errors a model is making, thereby illuminating its strengths and weaknesses.

B. Results of the experiment

The dataset was randomly split into a training set and a test set with 75% of the data allocated to the training set and 25% to the test set.

First we trained our models on the imbalanced dataset. Table I displays the experimental results for the imbalanced dataset. It is evident that Linear Discriminant Analysis (LDA) gives the highest F-1 score of 60% followed by Extra Trees Classifier (ET) which gives an F-1 score of 56%.

TABLE I. EXPERIMENTAL RESULTS ON THE IMBALANCED DATASET

Model	F-1 Score	Precision	Recall	Accuracy	MCC
ET	0.56	0.59	0.55	0.88	0.14
RF	0.48	0.85	0.50	0.91	0.06
LDA	0.60	0.67	0.58	0.90	0.24
SGD	0.54	0.71	0.53	0.91	0.16
LR	0.54	0.71	0.53	0.91	0.17

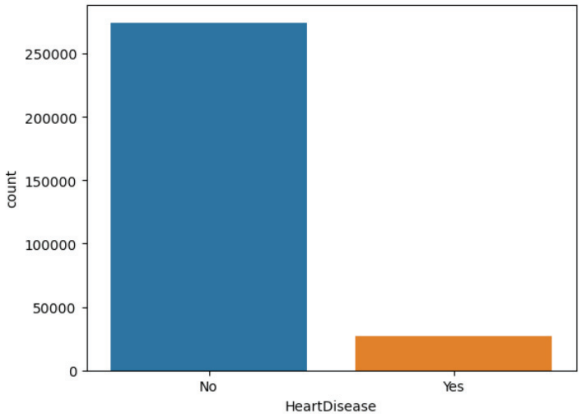

Fig. 3. Imbalanced dataset

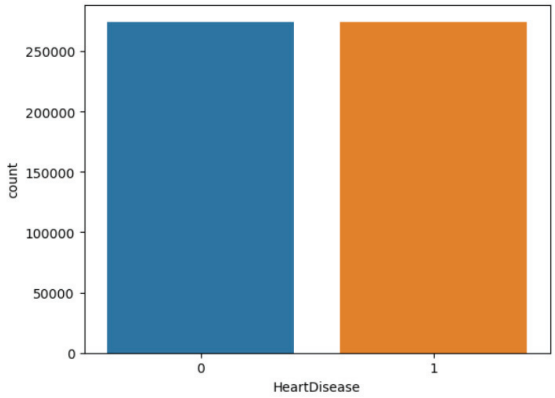

Fig. 4. Balanced dataset after oversampling

Table II displays the experimental results for the random oversampling technique. We can see that Extra Trees Classifier (ET) gives the highest F-1 score of 95% followed by Random Forest Classifier which gives an F-1 score of 75%. We can see that oversampling has significantly increased the F-1 score.

TABLE II. EXPERIMENTAL RESULTS ON THE BALANCED DATASET USING RANDOM OVERSAMPLING

Model	F-1 Score	Precision	Recall	Accuracy	MCC
ET	0.95	0.95	0.95	0.95	0.91
RF	0.75	0.75	0.75	0.75	0.51
LDA	0.74	0.74	0.74	0.74	0.49
SGD	0.74	0.74	0.74	0.74	0.49
LR	0.74	0.74	0.74	0.74	0.49

Table III displays the experimental results of the SMOTE technique. We can see that Extra Trees Classifier gives the highest F-1 score of 81%. LDA, SGD and LR all gave an F-1 score of 75%. Random Forest Classifier (RF) did not perform well with SMOTE oversampling with an F-1 score of only 71%.

TABLE III. EXPERIMENTAL RESULTS ON THE BALANCED DATASET USING SMOTE

Model	F-1 Score	Precision	Recall	Accuracy	MCC
ET	0.81	0.84	0.81	0.81	0.66
RF	0.71	0.78	0.72	0.72	0.50
LDA	0.75	0.75	0.75	0.75	0.50
SGD	0.75	0.75	0.75	0.75	0.51
LR	0.75	0.74	0.75	075	0.50

Table IV shows the experimental results of the ADASYN technique. Random Forest Classifier (RF) performed the best using ADASYN with an F-1 score of 75%. LDA, SGD and LR give an F-1 score of 74%. ET did not perform well with ADASYN with an F-1 score of only 69%.

TABLE IV. EXPERIMENTAL RESULTS ON THE BALANCED DATASET USING ADASYN

Model	F-1 Score	Precision	Recall	Accuracy	MCC
ET	0.69	0.79	0.71	0.71	0.50
RF	0.75	0.78	0.75	0.75	0.54
LDA	0.74	0.74	0.74	0.74	0.48
SGD	0.74	0.74	0.74	0.74	0.49
LR	0.74	0.74	0.74	0.74	0.48

V. CONCLUSION AND FUTURE WORKS

Given the dataset's class imbalance, prediction of CHD can be challenging. However, this challenge can be effectively managed, and the performance of CHD prediction models can be improved by employing a variety of oversampling methods. In imbalance data almost all the algorithms show 90% accuracy but only LDA shows a F-1 score of 60%. Oversampling techniques have improved the F-1 scores of all the algorithms. Highest F-1 score achieved is 95% by Extra Trees Classifier (ET) using random oversampling. ET has performed well using all oversampling techniques. In our study, we showed that using oversampling techniques can significantly enhance the performance of the CHD prediction model by tackling multiple aspects of the class imbalance issue. Our strategy involves balancing the dataset and improving the F-1 score of the predictions. The proposed approach can be applied to other areas of healthcare that face similar issues with class imbalance. In conclusion, our study makes an important impact to the fields of healthcare and machine learning, and it can help medical experts make better decisions about the treatment and diagnosis of coronary heart disease.

Experimenting with deep learning models like CNNs and RNNs is a future possibility [25]. Feature selection and dimensionality reduction have the potential to improve analysis in this sector. Using hybrid and ensemble techniques can improve F-1 score of the predictions.

REFERENCES

[1] Cardiovascular diseases (cvds), May 2023, [online] Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases(cvds).

[2] C. Shao, J. Wang, J. Tian and Y.-D. Tang, "Coronary artery disease: from mechanism to clinical practice", Coronary Artery Disease: Therapeutics and Drug Discovery, pp. 1-36, 2020.

[3] Kwakye, K. and Dadzie, E., "Machine Learning-Based Classification Algorithms for the Prediction of Coronary Heart Diseases, ", arXiv preprint arXiv:2112.015, 2021.

[4] Chakraborty, Srijita, et al. "Predicting Diabetes: A Comparative Study of Machine Learning Models." 2023 OITS International Conference on Information Technology (OCIT), pp. 743–48, 2023.

[5] Panda, Amiya Ranjan, et al. "System for Detecting Drowsiness in Drivers." 2023 OITS International Conference on Information Technology (OCIT), pp. 738–342, 2023.

[6] M. A. Mahmud Pranto, N. Al Asad, M. I. Adnan Palash, A. M. Islam and M. Shamim Kaiser, "Covid-19 chest x-ray classification with augmented gan", Proceedings of International Conference on Fourth Industrial Revolution and Beyond 2021, pp. 125-139, 2022.

[7] Banerjee, Amitayas, et al. "Comparative Analysis of Machine Learning and ANN Models for Mortality Prediction in RTAs." 2023 OITS International Conference on Information Technology (OCIT), pp. 698–702, 2023.

[8] S. Patidar, A. Jain and A. Gupta, "Comparative analysis of machine learning algorithms for heart disease predictions", 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 1340-1344, 2022.

[9] Bharti, R., Khamparia, A., Shabaz, M., Dhiman, G., Pande, S. and Singh, P., "Prediction of Heart Disease using a Combination of Machine Learning and Deep Learning," Computational intelligence and neuroscience, 2021.

[10] S. V. Jinny and Y. V. Mate, "Early prediction model for coronary heart disease using genetic algorithms hyper-parameter optimization and machine learning techniques", Health and Technology, vol. 11, pp. 63-73, 2021.

[11] P. Motarwar, A. Duraphe, G. Suganya and M. Premalatha, "Cognitive Approach for Heart Disease Prediction using Machine Learning," 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), pp. 1-5, 2020.

[12] S. K. J. and G. S., "Prediction of Heart Disease Using Machine Learning Algorithms," 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT), pp. 1-5, 2019.

[13] R. YILMAZ and F. H. YAĞIN, "Early detection of coronary heart disease based on machine learning methods", Medical Records, vol. 4, no. 1, pp. 1-6, 2022.

[14] S. Farzana and D. Veeraiah, "Dynamic Heart Disease Prediction using Multi-Machine Learning Techniques," 2020 5th International Conference on Computing, Communication and Security (ICCCS), pp. 1-5, 2020.

[15] A. N. V. K. Swarupa, V. H. Sree, S. Nookambika, Y. K. S. Kishore and U. R. Teja, "Disease Prediction: Smart Disease Prediction System using Random Forest Algorithm," 2021 IEEE International Conference on Intelligent Systems, Smart and Green Technologies (ICISSGT), Visakhapatnam, India, pp. 48-51, 2021.

[16] B. Dhananjay, N. P. Venkatesh, A. Bhardwaj and J. Sivaraman, "Cardiac signals classification based on Extra Trees model," 2021 8th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, pp. 402-406, 2021.

[17] R. Bhuvana, S. Maheshwari and S. Sasikala, "Predict the Heart Disease Using a Logistic Regression Classifier Algorithm," 2023 12th International Conference on System Modeling & Advancement in Research Trends (SMART), Moradabad, India, pp. 649-652, 2023.

[18] H. Padalko, K. Bazilevych and M. Butkevych, "Heart Failure Development Prediction using Stochastic Gradient Descent Optimization," 2022 IEEE 9th International Conference on Problems of Infocommunications, Science and Technology (PIC S&T), Kharkiv, Ukraine, pp. 297-300, 2022.

[19] F. Kamalov, H. -H. Leung and A. K. Cherukuri, "Keep it simple: random oversampling for imbalanced data," 2023 Advances in Science and Engineering Technology International Conferences (ASET), Dubai, United Arab Emirates, pp. 1-4, 2023.

[20] I. Dey and V. Pratap, "A Comparative Study of SMOTE, Borderline-SMOTE, and ADASYN Oversampling Techniques using Different Classifiers," 2023 3rd International Conference on Smart Data Intelligence (ICSMDI), Trichy, India, pp. 294-302, 2023.

[21] C. Lu, S. Lin, X. Liu and H. Shi, "Telecom Fraud Identification Based on ADASYN and Random Forest," 2020 5th International Conference on Computer and Communication Systems (ICCCS), Shanghai, China, pp. 447-452, 2020.

[22] P. V. Sairam and L. K, "Automatic Stock Market Prediction using Novel Long Short Term Memory Algorithm compared with Logistic Regression for improved F1 score," 2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM), Gautam Buddha Nagar, India, pp. 578-582, 2022.

[23] V. L. Boiculese, G. Dimitriu and M. Moscalu, "Improving recall of k-nearest neighbor algorithm for classes of uneven size," 2013 E-Health and Bioengineering Conference (EHB), Iasi, Romania, pp. 1-4, 2013.

[24] C. Gao, Y. Zhang, D. Lo, Y. Shi and J. Huang, "Improving the Machine Learning Prediction Accuracy with Clustering Discretization," 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, pp. 0513-0517, 2022.

[25] Tiwari, Abhinandan Kumar, et al. "Parametric Examination on Optimized Deep Learning Based Melanoma Detection." 2021 Fourth International Conference on Electrical, Computer and Communication Technologies (ICECCT), pp. 1–8, 2021.