[,*] Shivganga Udhan, [1]
[2]Bankat Patil,
[3]Suvarna Patil,
[4]Sneha Kanawade,
[5]Priyadarshani Doke,
[6]Deepali Hajare

Machine intelligence based early prediction methods for Cardiovascular Diseases (CVDs)

Abstract: - Exploring early indications of heart disease is challenging in today's society. It may lead to death if not detected early. Early identification of cardiac disease in remote, rural, and semi-urban locations in developing nations can be greatly aided by an accurate decision support system (DSS). To assist in heart-related disease detection at an early stage, this system offers a DSS which uses artificial learning techniques using the patient's clinical details. A hybridized class balancing methodology using the chi2 feature selection algorithm, random oversampling, and undersampling techniques were used to find the suitable features from the presented dataset. Standard scalar approaches have also been utilized for data preprocessing. In the last stage of developing the suggested Artificial Intelligence (AI) system, the system used support vector machines (SVM), naive Bayes, random forest classifiers, logistic regression, and improved Artificial Neural Network (ANN) classifiers. Tests were conducted on the Python-based simulation environment to evaluate the proposed system. Evaluation of the system is completed using the Cleveland heart disease dataset by the machine repository at UCI. A 96.74% accuracy rate was reached, which is better than some previously published models for heart-related disease prediction.

Keywords: Machine Learning, Prediction, cardiovascular disease, ANN, feature extraction.

I. INTRODUCTION

Cardiovascular disease (CVD) has long been considered as the most serious and deadly of all human diseases. In middle and old life, males are more likely to develop CVD than women [1], with similar health issues detected in children [2][3]. Healthcare systems around the world are under strain and in risk as a result of growing cardiovascular disorders and their related high death rates.

As per WHO report, the heart-related disease accounts for 33 percent of all deaths worldwide. CVD is predicted to affect 18.9 million lives in 2022, or 32% of all fatalities worldwide. A heart attack or stroke killed 85 % of these people. [4][5]. According to research, more than 26 million people worldwide are affected by CVD, with more than 3 million new patients being detected yearly. CVD kills half of all patients within two years after diagnosis, and around 3% of total healthcare spending is spent on treating it [6]. Multiple tests are required to accurately predict cardiovascular disease. The medical staff's lack of competency may result in inaccurate forecasts. [7].

Early diagnosis can be difficult [8]. It can be difficult to treat heart disease surgically, especially in developing countries where skilled medical workers and diagnostic equipment are scarce [9]. Authors have compared machine learning methods for predicting Coronary Artery Diseases by constructing pooled area curves (PUC) [10]. Researchers have predicted diabetes and heart diseases using Machine Learning models [11].

The public can compare prediction models using heart disease datasets, which are available online. Data from large databases can be utilised to develop the best prediction model using artificial intelligence and machine learning. The need to reduce mortality from CVD has been stressed in recent research looking at adults and children. Although the available medical datasets are unreliable and redundant, preparation is required [12]. The accuracy of prediction models is enhanced by selecting the most relevant features. It is critical to use the appropriate ML techniques to create reliable forecasting system [13].

If risk variables satisfy the three requirements of significant independent influence on heart disease, high prevalence, and controllability or treatability, then it is essential to consider whether they have the potential to lower risks. [14]. Various risk variables and features have been incorporated into the models of CVD predictors by various studies. Features employed in the construction of CVD prediction models include sex, age, and chest pain (cp); higher FBS is associated with diabetes [15][16].

[1,*]Shivganga Udhan, *Dr. Babasaheb Ambedkar Marathwada University. Aurangabad, India*
[2]Bankat Patil, *Dr. Babasaheb Ambedkar Marathwada University. Aurangabad, India*
[3]Suvarna Patil, *Dr. D.Y.Patil Institute of Engineering Management and Research Akurdi*
[4]Sneha Kanawade, *Dr. D.Y.Patil Institute of Engineering Management and Research Akurdi*
[5]Priyadarshani Doke, *Dr. D.Y.Patil Institute of Engineering Management and Research Akurdi*
[6]Deepali Hajare, *Dr. D.Y.Patil Institute of Engineering Management and Research Akurdi*
[1]shivganga168@gmail.com, [2]patilbankat@gmail.com [3]suvarnapat@gmail.com, [4]sneha.kanawade@gmail.com,
[5]priyadarshani.doke@dypiemr.ac.in, [6]deepali.hajare@dypiemr.ac.in

Prediction accuracy and reliability can only be achieved if a minimum of 14 features are present. [17]. Currently, researchers are having difficulty combining these features with the necessary ML algorithms for the prediction of cardiovascular illness. [18]. Machine learning algorithms fit best when trained on appropriate datasets [19][20]. Data preparation methods like ref selection, LASSO, and data mining can help in the preparation of the data for more accurate prediction. However, the algorithms depend on the consistency of the train and test data. It is possible to estimate the likelihood that a disease would manifest itself using hybrid classifier. Researchers used various methods to construct classifiers and hybrid models [12] [17].

Heart disease prediction may be hampered by a lack of sufficient medical datasets, feature selection, machine learning algorithms, and in-depth analysis, just a few of the roadblocks. Our research strives to address the knowledge gaps in order to develop a comprehensive CVD prediction model.

I.1. Contributions of the paper

1. Using rank values, the Chi2 K Best feature selection algorithm extracts the critical factors from medical sources. This method can help to solve the problems of overfitting and underfitting in machine learning.

2. The mechanism of random oversampling and standard scaling was implemented to achieve the best results using various classifiers. Several types of supervised models used in this research, such as Decision Trees, AdaBoost, K-Nearest Neighbor, Random Forest, Gradient Boosting, as well as an ANN classifier.

I.2. Organization of the paper

The structure of this article is described as: Second Section gives the scope and the aim. Section III of this study represents an overview of pertinent studies on classifiers and hybrid approaches for heart-related disease prediction. The fourth section shows the recommended procedures along with processes. Section IV discusses the preparation of data, preprocessing, various machine learning and ANN methods, feature selection, and random oversampling and undersampling. In Section V, the outcomes of the system's implementation are discussed. Section V contains a comprehensive discussion of the data's statistical significance, runtime and computational complexity, and performance analysis. Section VI includes some conclusions and suggestions for further research.

II. RELATED WORK

Different AI methods, such as ML and DL, have significantly impacted several industries by enabling more precise predictions and analysis across a variety of system domains. The ability to reliably predict the start of numerous diseases in advance has become crucial in medical diagnosis, helping both doctors and patients. A range of ML and DL-based DSS is suggested by the researchers in CVD management for heart-related disease prediction.

An ML-based DSS for cardiac disease prediction was proposed by Pooja Rani et al. [21]. An equation chaining method utilizing multivariate imputation was used by the authors for missing value filling. The Cleveland heart disease patient dataset was utilized to identify pertinent features using a selection technique hybrid feature that combines the Genetic Algorithm in which feature elimination is applied recursively. Also, for data preparation, scalar techniques and Synthetic Minority Oversampling Technique (SMOTE) is used. The last phases of hybrid system includes Naive Bayes, logistic regression, SVM, randomly-generated forests (RF), and Adaboost classifiers (ADC). The RF classifier was determined to have the highest accuracy using the method, with a rate of 86.6%.

Minimize false alarms, minimize process costs, and maximize the accuracy of label predictions, F.A. Mohammad et al. [22] proposed a model with feature optimization. It was determined how closely the feature and decision-label correlations are connected by utilizing the feature n-gram sequence data (positive, negative). The proposed model gives feature optimization by discrete weights (FODW), and evaluation is done using performance parameters viz precision, specificity, sensitivity, accuracy, and the f-measure to compare it to other recent models like HRFLM and HIFS.

The Cleveland Heart Sample database was used by G. Magesh and P. Swarnlatha to develop a cluster-based Decision Tree learning (CDTL) system to predict heart-related diseases. [23]. When it comes to dividing up the original collection, it was done by the target label. The class combination was produced using data with high dispersion. The features related to an entropy-based partition are crucial. Finally, RF performance and all elements of heart disease prediction are made relevant. By using the CDTL method, the prediction accuracy is increased from 76.70 to 89.30 % (without CDTL).

G.Sarayan and A. Pravin studied classification techniques in the Machine Learning model to choose the best attributes. [24]. More sensitive qualities were selected for consideration in the second phase of the experiment utilizing the PSO method. The author was able to effectively identify cardiac illness with 90% accuracy and 94%

sensitivity utilizing the suggested PSO-GSA (Particle swarm optimization with a globally sensitive approach) using machine learning classification methods.

In a comprehensive evaluation of ML-based heart disease prediction, A. Kondababu et al. [25] investigated various classification methods for the prediction of heart-related disease. Several existing methods were investigated, and the Hybrid Random Forest with a Linear Model (HRFLM) strategy was shown to have an accuracy level of up to 88.7%. A composite dataset (Cleveland, Switzerland, Long Beach, VA, Stat log, and Hungarian) was used in the study by P. Ghosh et al. [26]. This was done using the Least Absolute Shrinkage Selection Operator (LASSO) method. For building hybridized classifiers, traditional classifiers were mated together with bagging and boosting techniques. The suggested model has the best accuracy of 99.05% using approaches for determining the best features and the Random Forest Bagging Method.

Bhanu Prakash Doppala et al. [27] improved the identification of coronary artery disease by combining a genetic algorithm and radial basis function (GA-RBF). After reducing the number of characteristics from 14 to 9, the suggested system performed substantially better, with an increase in accuracy from 85.40% to 94.20%.

In their study, F. Ali et al. proposed an ensemble Deep Learning and a feature fusion-based system for predicting heart disease [28]. Sensor data was used to obtain healthcare data. Using the information-gathering process, superfluous and redundant features were removed. It was then utilized to classify heart disease using the ensemble DL model with accuracy of 98.50%.

Based on essential feature scores, Armin Yazdani et al. used weighted associative rule mining to predict cardiac disease. [29]. A set of feature ratings and criteria used in the heart disease diagnosis were examined for validity by cardiologists. With a 98% degree of confidence, scientists were able to make a prediction of heart disease using the UCI dataset, which is frequently utilized in heart disease research.

Pre-trained DNN is proposed for feature extraction, dimensionality reduction is done using Principal Component Analysis and Logistic Regression (LR) is used for predicting the heart disease [33]. In [34], Deep Neural Network (DNN) Model with four hidden layers is used for detecting coronary heart diseases. The most successful model is given using various combinations of three layers: input layer, hidden layer, and output layer. A systematic review of all the methods used for deart disease detection is done by authors [35]. Comparative analysis of performance of all models is given by authors [36][37].

The comparative analysis of different study for different selection methods is shown in Table 1.

TABLE 1. Comparative Analysis of Proposed System with Existing System

Study	Publication Year	Feature Selection Method	Classification model
Verma and Mathur	2019	Cuckoo search	Multilayer perceptron
Jabbar et al.	2016	Chi-Square	RF
Latha and Jeeva	2019	Feature subset	NB, RF, MLP
Tama et al	2020	PSO	Ensemble method
Rani et al.	2021	GA and RFE	RF

III. PROPOSED SYSTEM MODEL

3.1 Methodology

A DSS that is based on an ANN for the forecasting of cardiac disease has been proposed in this study. The three steps of the proposed system are data collecting, the preparation of the data, and model creation. The preprocessing stage involves the selection of features, feature scaling, and class balancing. Then, a Chi-square K Best feature selection approach is employed to make the feature selection. Using a standard scalar, the coefficient value of each feature is equalized, guaranteeing that each feature has a SD of 1 and a mean of 0. There are 165 cases in the dataset that correspond to class 0, and there are 138 examples that belong to class 1. Random oversampling and undersampling methods are used to balance classes [31]. Using LR, RF, NB, SVM, SGD, KNN, AdaBoost classifier, and ANN, classification is applied on a subset of selected features. The classifier also forecasts if someone has heart disease or not.

Algorithm1: Model without pre-processing

Here different ML algorithms are employed on heart disease dataset and their comparison is done.
1. Collect heart disease dataset
2. EDA to understand dataset and list some observations.
3. Train test splitting
4. Model creation using different classification methods like Logistics Regression, Naive Bayes, SVM, KNN, RF, SGD, XGB
5.Compare performance of all models with ideal condition by drawing ROC curve.

Algorithm2: Model with Class Balancing
Here Imbalanced data is handled using random oversampling and random under sampling methods. Results obtained are better than the results of Algorithm1
1. Collect heart disease dataset
2.EDA to understand dataset and list some observations.
3. Preparation of the data
a. Handling imbalanced dataset
i. Random Over sampling
ii. Random under sampling
4. Train Test splitting
5. Model creation using different classification methods like Logistics Regression, Naive Bayes, SVM, KNN, RF, SGD, XGB
6.Compare performance of all models

Algorithm3: Model with feature selection
In this algorithm top ten features are identified using Chi Square K best feature selection technique in combination with class balancing techniques like random over sampling and random under sampling.
1. Collect heart disease dataset
2. EDA to understand dataset and list some observations.
3. Preparation of the data
a. Feature selection using Chi2 K best feature selection approach
b. Handling imbalanced dataset
i. Random Over sampling
ii. Random under sampling
c. Using a standard scalar, the coefficient value of each feature is equalised, guaranteeing that each feature has a SD of 1 and a mean of 0.
4. Train Test splitting
5. Model creation using different classification methods like Logistics Regression, Naive Bayes, SVM, KNN, RF, SGD, XGB
6. Compare performance of all models

3.2 Dataset

The research utilized data from the UCI library on Cleveland Heart Disease (CHD). The dataset consists of 14 features, with eight being categorical and the remaining six being numeric.
Table 2 provides examples of the dataset's features along with their corresponding descriptions [21]

TABLE 2. Feature details

Feature Code: Feature name	**Description**
AG: Age	Age between 29 to 77 years
SX: Sex	Female: 0, Male: 1
CP: Type of chest pain	Asymptomatic: 4, non-angina pain: 3, atypical angina: 2, Typical angina: 1
RBP: Resting blood pressure	Between 94 mm Hg to 200 mm Hg
SCHOL: Cholesterol	Between 126 mg/dl to 564 mg/dl
FABS: Fasting blood sugar	FBSR > 120 mg/dl (false: 0, true:1)

REAR: Resting electrocardiographic results	Hypertrophy: 2, ST-T wave abnormality: 1, Normal: 0
HR: Maximum heart rate achieved	Between 71 to 202
EIAG: Exercise-induced angina	NO:0, YES:1
STD: ST depression induced by exercise relative to rest	Up sloping: 1, Flat: 2, down sloping: 3
SPE: The slope of the peak exercise ST segment	Between 0 to 6.2
NMVCF: Number of significant vessels (0–3) colored by fluoroscopy	Between 0 to 3
THALM: Thallium	reversible defect: 7, fixed defect: 6, Normal: 3
TARG: Target	Heart disease absent: 0, heart disease present: 1

3.3 Techniques for feature selection

Feature selection methods are crucial for the ML process since they allow the extraction of the finest attributes for categorization. Additionally, this shortens the execution time. The Chi2 feature selection method has been chosen. The chi-square test assists in solving feature selection issues by looking at the correlations between the features.

In statistics, the chi-square test is employed to determine the independence of two events. The observed count (O) and predicted count (E) for two variables can be derived from the data. The Chi-Square test calculates the difference values between the observed count (O) and the expected count (E), as depicted in Equation (1).

$$x_f^2 = \sum \frac{(Ov_i - Ev_i)^2}{Ev_i} \qquad (1)$$

Here:
f is a degree of freedom
Ov is observed value(s)
Ev is expected value(s)

3.4 Random oversampling & und oversampling

The bias in the training dataset can have an impact on many ML algorithms, leading to some of them ignoring the minority class. Because the minority class is given the most weighted projections, this is a cause for concern. One approach to addressing the class imbalance issue is random resampling of training dataset. There are two basic ways of randomly resampling an unbalanced dataset, one termed under-sampling and the other called oversampling. In this study, the dataset we are dealing with is skewed. There are a total of 303 datasets accessible, of which 165 pertain to heart defects and 138 to healthy hearts. Figure 1 shows how the skewed data is analyzed.

IV. RESULTS AND DISCUSSION

To evaluate the results, the experimental setup is performed using the Machine Learning models from Python libraries like sklearn, pandas, numpy. Other libraries like matplotlib and seaborn are used for anlyzing the results.

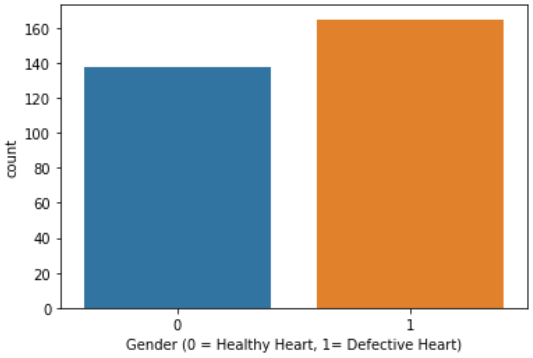

Figure 1. Imbalance Cleveland Dataset

The findings are compared with the existing system and elaborated in this section. Random Forest, logistic regression, Adaboost, SVM, and Naive Bayes classification methods used to identify patients with heart disease. Several medical factors used to identify heart-related disease in the dataset. For categorization, these factors were employed, with class 1 signifying illness and class 0 denoting disease-free status. Precision, accuracy, specificity, sensitivity, and F-measure were used for evaluating the systems performance.

4.1 Classification algorithm performance on the imbalanced dataset and with all features

Initially, all features of the dataset tested without preprocessing, data balancing, or feature selection. Table 3 illustrates the classifier's performance throughout the entire feature set. The NB classifier's overall performance is good as it can handle continuous and discrete data with high scalability, whereas the SGD classifier performs worst.

TABLE 3. classifier's performance for all features.

Classification Model	Precision	Recall	Accuracy	F1-Score	Training_time (ms)	Testing_time (ms)
Logistic Regression (LR)	0.80	0.85	82.25	0.82	100.71	2.12
Naïve Byes (NB)	0.86	0.93	83.61	0.89	5.72	3.61
Support Vector Machine (SVM)	0.81	0.91	81.97	0.85	364.32	3.79
K-Nearest Neighbour (KNN)	0.55	0.57	65.57	0.56	3.40	8.57
Random Forest (RF)	0.75	0.84	81.97	0.80	16.65	2.96
XG-Boost	0.78	0.84	73.77	0.81	3.73	1.33
Stochastic Gradient Descent (SGD)	1.00	0.46	63.93	0.63	63.23	2.21
ANN	0.79	0.83	72.73	0.80	1669.74	3.85

4.2 Classification algorithm performance with Random Oversampling and feature selection.

After using the scaling preprocessing procedure, we looked at the classifiers' results once again. The input dataset was subjected to the usual scalar method. We used several data balancing approaches, such as Radom oversampling (Figure 2) and undersampling(Figure 5), to correct the imbalance in the Cleveland dataset. Classifier methods were used for a balanced dataset in this part. We also used the Chi2 K Best selection method for feature selection. We selected the top 10 features. Figure 4 represents the top 10 features selected using the Chi2 K Best selection method. The accuracy of SVM decreased from 81.97 % to 79.00%, whereas the accuracy of RF increased by 0.03%. A massive increase in ANN accuracy is found, and it has increased from 72.73% to 95.63%. The results show that scaling has a highly positive effect on ANN, RF, and NB classifier performance but not on SVM, XGBoost, or SGD classifier performance. Comparative performance analysis of the effects of scaling, random oversampling (class balancing), and feature selection approaches for classifiers is shown in Table 4, and its graphical comparison is shown in Figure 3.

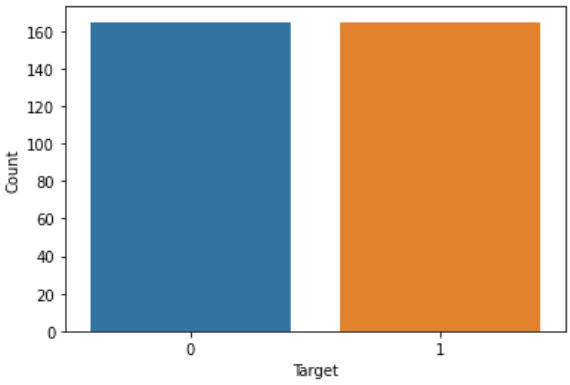

Figure 2. Random Oversampling class balanced data

TABLE 4. Performance Analysis of Classification Model After Random Oversampling

Classification Model	Accuracy	Precision	Recall	F1-Score	Training time (ms)	Testing time (ms)	False alarm rate (%)
Logistic Regression (LR)	84.00	0.90	0.80	0.85	62.39	0.22	07
Naïve Byes (NB)	85.25	0.80	0.84	0.82	4.13	0.44	14
Support Vector Machine (SVM)	79.00	0.84	0.77	0.81	153.19	1.87	14
K-Nearest Neighbour (KNN)	64.00	0.70	0.66	0.68	1.18	6.48	16
Random Forest (RF)	82.00	0.85	0.83	0.84	14.78	1.65	13
XG-Boost	82.00	0.88	0.80	0.84	2.48	0.15	02
Stochastic Gradient Descent (SGD)	61.00	0.61	0.89	0.72	54.72	1.16	14
ANN	95.63	0.89	0.81	0.82	1616.70	4.46	01

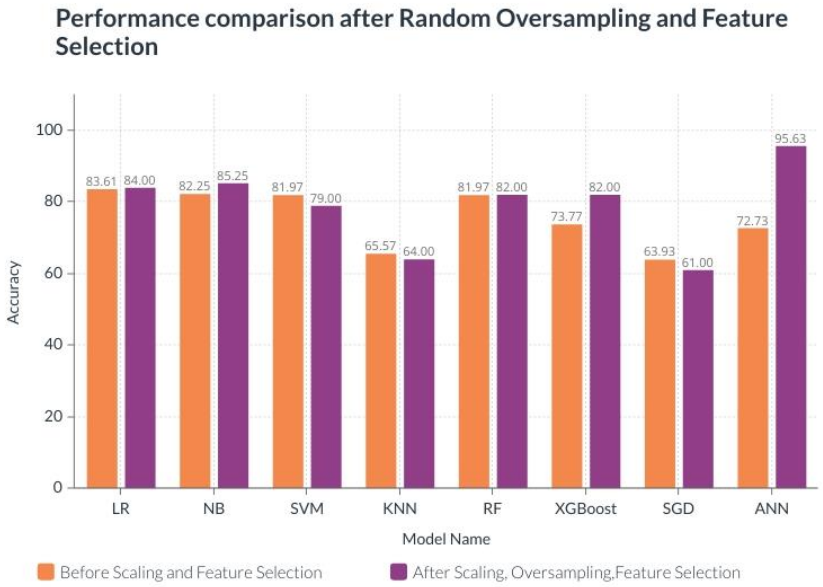

Figure 3. Performance comparison after Random Oversampling and Feature Selection

4.3 Classification algorithm performance with Random Undersampling and Feature selection

In this section, classifier algorithms are applied to the Random Under sampling balanced dataset. Further, the feature selection technique using the Chi2 K Best selection algorithm. is applied. Then the top 10 features are selected after applying the Chi2 K Best Selection algorithm. SVM accuracy improved by 2.03%, whereas LR accuracy improved by 0.44%. The accuracy of NB had changed from 82.25% to 85.25%, an increase of 3.00%, whereas KNN and RF algorithms had shown a decline in accuracy after random under-sampling.

The ultimate model we applied, i.e., ANN, showed an increase in accuracy from 72.73% to 96.74 %. Though the training time required for this model is more testing, the time required is significantly less.

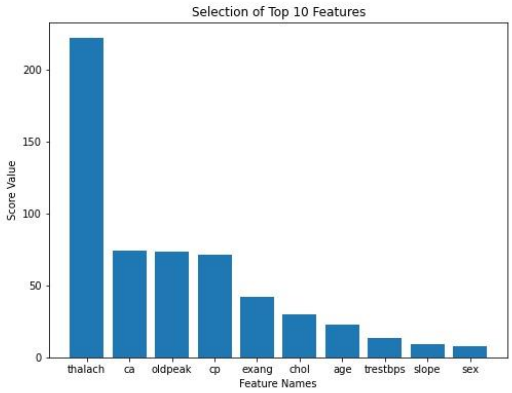

Figure 4. Top 10 K Best Selected Features

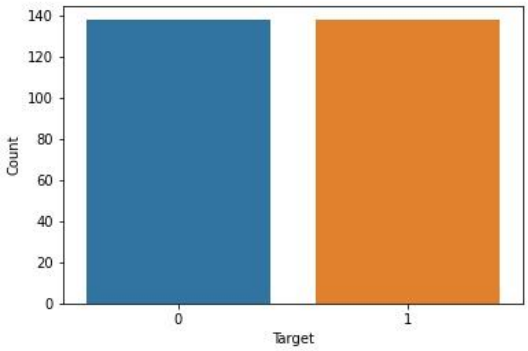

Figure 5. Random undersampling class balanced data

Results indicate that scaling, class balancing using random undersampling, and feature selection have a very positive impact on the ANN algorithm. We achieved the best of all accuracy using this mechanism.

Table 5 shows a comparison of the effects of scaling, random undersampling (class balancing), and feature selection approach, and its graphical representation is shown in Figure 6.

TABLE *5.* Performance of Model after Random Undersampling

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**Training time (ms)**	**Testing time (ms)**	**False alarm rate (%)**
Logistic Regression (LR)	84.00	0.83	0.77	0.80	60.22	0.25	07
Naïve Byes (NB)	85.25	0.80	0.84	0.82	4.39	0.40	14
Support Vector Machine (SVM)	84.00	0.84	0.89	0.76	1512.28	1.62	15
K-Nearest Neighbour (KNN)	61.00	0.66	0.66	0.66	1.02	6.11	16
Random Forest (RF)	79.00	0.82	0.80	0.81	13.12	0.96	13
XG-Boost	79.00	0.81	0.83	0.82	2.01	0.10	02
Stochastic Gradient Descent (SGD)	59.00	0.51	0.96	0.67	52.92	0.97	14
ANN	96.74	0.89	0.90	0.89	1558.79	4.31	01

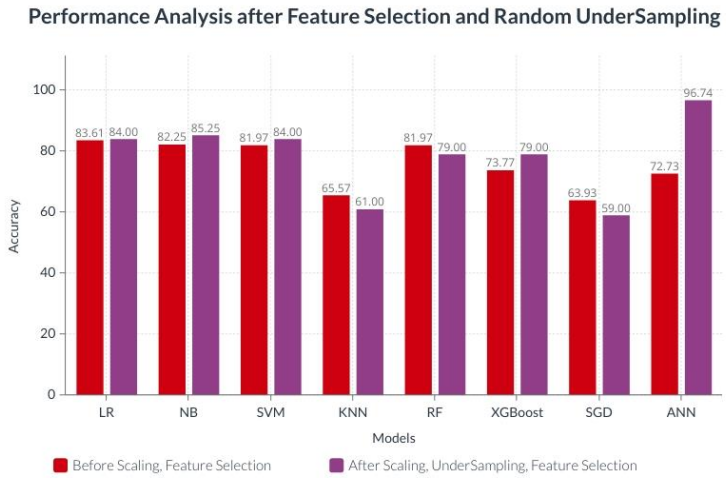

Figure 6. Performance comparison after Random undersampling and Feature Selection

4.4 Performance analysis after feature selection

Using feature selection, classifier accuracy was increased even more. For feature selection, the Chi2 K Best selection method was used. From existing features in the given dataset, the top 10 features are chosen using this feature selection mechanism. Table 6 demonstrates how feature selection enhanced the effectiveness of the ANN, XGBoost, and Random Forest classifiers. In the case of ANN, the performance had increased from 72.73% to 96.74%. In XGBoost, the performance increased from 73.7 to 81.97 %, and in the case of Random Forest, it increased from 81.97% to 85.25%. NB, SVM accuracy had decreased the accuracy performance from 82.25% to 67.21% and from 81.97% to 77.05% in SVM. In the case of LR, the performance of accuracy decreased from 83.61 to 80.33%.

TABLE 6. Final Performance analysis before and after data balancing, feature selection

	Without Data Balancing &	**Data Balancing Feature Selection**	
		Oversampling	**Undersampling**
	All Features	**Top 10 Features**	**Top 10**
Model	Precision	Recall	Accuracy
Logistic Regression (LR)	83.61	78.63	80.33
Naïve Byes (NB)	82.25	42.62	67.21
Support Vector Machine	81.97	78.69	77.05
K-Nearest Neighbour (KNN)	65.57	59.02	62.30
Random Forest (RF)	63.93	42.62	67.21
XG-Boost	81.97	80.33	85.25
Stochastic Gradient Descent	73.70	80.33	81.97
ANN	72.73	95.63	96.74

Results show that the Chi2 K Best feature selection method with ten features and ANN that combined input, hidden, and output layers achieved the desired performance. Dimensionality reduction via feature selection has enhanced the efficiency of ANN. This not only reflected in the improvement in classification performance analysis but even the time required for training and testing of the model had been reduced.

V. CONCLUSION

Lack of early detection is the main reason for heart disease-related deaths. To mitigate the effects of the proposed research to develop the advanced system of decision assistance for heart-related disease as ANN. The system contribution is to develop an accurate DSS compared to existing approaches for detection at the initial stage of heart disease. The system implemented the process to find the optimal methods for feature selection using the class balancing method of random oversampling and undersampling and classifier choice and has incorporated those

findings into their hybrid DSS. The suggested system was tested and compared using the Cleveland dataset in a python-created simulated environment. It has performed better than any other hybrid decision assistance system that has been published 96.74% was the highest accuracy achieved by ANN systems.

In the future, the performance of the system can be improved by using an ant colony and particle swarm optimization approach. The current system can detect only CVD. Furthermore, the system can be improved for the diagnosis of other diseases.

REFERENCES

[1] Trevisan, C., Sergi, G., & Maggi, S. (2020). Gender differences in brain-heart connection. Brain and heart dynamics, 937-951.

[2] Ryu, H., Moon, J., & Jung, J. (2020). Sex differences in cardiovascular disease risk by socioeconomic status (SES) of workers using National health information database. International journal of environmental research and public health, 17(6), 2047.

[3] Jousilahti, P., Vartiainen, E., Tuomilehto, J., & Puska, P. (1999). Sex, age, cardiovascular risk factors, and coronary heart disease: a prospective follow-up study of 14 786 middle-aged men and women in Finland. Circulation, 99(9), 1165-1172.

[4] "Cardiovascular diseases (CVDs)." https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (accessed Dec. 26, 2021).

[5] Uyar, K., & İlhan, A. (2017). Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks. Procedia computer science, 120, 588-593.

[6] Haq, A. U., Li, J. P., Memon, M. H., Nazir, S., & Sun, R. (2018). A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. Mobile information systems, 2018, 1-21.

[7] Pouriyeh, S., Vahid, S., Sannino, G., De Pietro, G., Arabnia, H., & Gutierrez, J. (2017, July). A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease. In 2017 IEEE symposium on computers and communications (ISCC) (pp. 204-207). IEEE.

[8] Mourao-Miranda, J., Bokde, A. L., Born, C., Hampel, H., & Stetter, M. (2005). Classifying brain states and determining the discriminating activation patterns: support vector machine on functional MRI data. Neuroimage, 28(4), 980-995.

[9] Ghwanmeh, S., Mohammad, A., & Al-Ibrahim, A. (2013). Innovative artificial neural networks-based decision support system for heart diseases diagnosis.

[10] Chen, J. I. Z., & Hengjinda, P. (2021). Early prediction of coronary artery disease (CAD) by machine learning method-a comparative study. Journal of Artificial Intelligence, 3(01), 17-33.

[11] Dey, A., Chanda, P. B., & Sarkar, S. K. (2020, November). Patient Health Observation and Analysis with Machine Learning and IoT Based in Realtime Environment. In 2020 Fifth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN) (pp. 196-201). IEEE.

[12] Amin, M. S., Chiam, Y. K., & Varathan, K. D. (2019). Identification of significant features and data mining techniques in predicting heart disease. Telematics and Informatics, 36, 82-93.

[13] Kausar, N., Palaniappan, S., Samir, B. B., Abdullah, A., & Dey, N. (2016). Systematic analysis of applied data mining based optimization algorithms in clinical attribute extraction and classification for diagnosis of cardiac patients. Applications of Intelligent Optimization in Biology and Medicine: Current Trends and Open Problems, 217-231.

[14] Mackay, J., Mensah, G. A., & Greenlund, K. (2004). The atlas of heart disease and stroke. World Health Organization.

[15] M. M. Alam et al., "D-CARE: A Non-invasive Glucose Measuring Technique for Monitoring Diabetes Patients," pp. 443–453, 2020, DOI: 10.1007/978-981-13-7564-4_38.

[16] Ashraf, M., Ahmad, S. M., Ganai, N. A., Shah, R. A., Zaman, M., Khan, S. A., & Shah, A. A. (2021). Prediction of cardiovascular disease through cutting-edge deep learning technologies: an empirical study based on TENSORFLOW, PYTORCH and KERAS. In International Conference on Innovative Computing and Communications: Proceedings of ICICC 2020, Volume 1 (pp. 239-255). Springer Singapore.

[17] Yahaya, L., Oye, N. D., & Garba, E. J. (2020). A comprehensive review on heart disease prediction using data mining and machine learning techniques. American Journal of Artificial Intelligence, 4(1), 20-29.

[18] Shouman, M., Turner, T., & Stocker, R. (2013). Integrating clustering with different data mining techniques in the diagnosis of heart disease. J. Comput. Sci. Eng, 20(1), 1-10.

[19] Mienye, I. D., Sun, Y., & Wang, Z. (2020). An improved ensemble learning approach for the prediction of heart disease risk. Informatics in Medicine Unlocked, 20, 100402.

[20] Wang, H., Huang, Z., Zhang, D., Arief, J., Lyu, T., & Tian, J. (2020). Integrating co-clustering and interpretable machine learning for the prediction of intravenous immunoglobulin resistance in kawasaki disease. IEEE Access, 8, 97064-97071.

[21] Rani, P., Kumar, R., Ahmed, N. M. S., & Jain, A. (2021). A decision support system for heart disease prediction

based upon machine learning. Journal of Reliable Intelligent Environments, 7(3), 263-275.

[22] Al-Yarimi, F. A. M., Munassar, N. M. A., Bamashmos, M. H. M., & Ali, M. Y. S. (2021). Feature optimization by discrete weights for heart disease prediction using supervised learning. Soft Computing, 25, 1821-1831.

[23] Magesh, G., & Swarnalatha, P. (2021). Optimal feature selection through a cluster-based DT learning (CDTL) in heart disease prediction. Evolutionary intelligence, 14, 583-593.

[24] Saranya, G., & Pravin, A. (2022). Hybrid global sensitivity analysis based optimal attribute selection using classification techniques by machine learning algorithm. Wireless Personal Communications, 127(3), 2305-2324.

[25] Kondababu, A., Siddhartha, V., Kumar, B. B., & Penumutchi, B. (2021). A comparative study on machine learning based heart disease prediction., 2021, doi: 10.1016/j.matpr.2021.01.475.

[26] Kondababu, A., Siddhartha, V., Kumar, B. B., & Penumutchi, B. (2021). WITHDRAWN: A comparative study on machine learning based heart disease prediction.

[27] Doppala, B. P., Bhattacharyya, D., Chakkravarthy, M., & Kim, T. H. (2021). A hybrid machine learning approach to identify coronary diseases using feature selection mechanism on heart disease dataset. Distributed and Parallel Databases, 1-20.

[28] Ali, F., El-Sappagh, S., Islam, S. R., Kwak, D., Ali, A., Imran, M., & Kwak, K. S. (2020). A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. Information Fusion, 63, 208-222.

[29] A. Yazdani, K. D. Varathan, Y. K. Chiam, A. W. Malik, and Yazdani, A., Varathan, K. D., Chiam, Y. K., Malik, A. W., & Wan Ahmad, W. A. (2021). A novel approach for heart disease prediction using strength scores with significant predictors. BMC medical informatics and decision making, 21(1), 194.

[30] Thanga Selvi, R., & Muthulakshmi, I. (2020). An optimal artificial neural network based big data application for heart disease diagnosis and classification model. J Ambient Intell Human Comput.

[31] Sáez, J. A., Krawczyk, B., & Woźniak, M. (2016). Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets. Pattern Recognition, 57, 164-178.

[32] "UCI Machine Learning Repository: Heart Disease Data Set." https://archive.ics.uci.edu/ml/datasets/heart+disease

[33] Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (1988). UCI machine learning repository-heart disease data set. School Inf. Comput. Sci., Univ. California, Irvine, CA, USA.

[34] Hassan, D., Hussein, H. I., & Hassan, M. M. (2023). Heart disease prediction based on pre-trained deep neural networks combined with principal component analysis. Biomedical Signal Processing and Control, 79, 104019.

[35] Udhan, S., & Patil, B. (2021). A systematic review of Machine learning techniques for Heart disease prediction. International Journal of Next-Generation Computing, 12(2).

[36] Udhan, S., & Patil, B. (2024). Cutting-Edge Neural Network for Early Cardiovascular Disease Prevention. International Journal of Intelligent Systems and Applications in Engineering, 12(3s), 05-16.

[37] Udhan, S., & Patil, B. (2023). Novel Deep Neural Network for Early Prediction and Prevention of Cardiovascular Disease.