A Comparative Performance Analysis of Machine Learning Models for Cardiovascular Disease Prediction Using Hyper-Parameter Tuning

Ms. Maitri Bhavsar[1],Dr. Manish Patel[2]

Assistant Professor,AI&DS, Sankalchand Patel College of Engineering, SPU,Visnagar,India[1]
Professor,AI&DS, Sankalchand Patel College of Engineering, SPU, Visnagar, India[2]
mhbai_spce@spu.ac.in[1], mmpatelit_spce@spu.ac.in[2]

Abstract: Cardiovascular disease (CVD) remains one of the leading causes of mortality worldwide, demanding reliable and early prediction systems. This study proposes an optimized machine learning framework that integrates advanced data pre-processing, dimensionality reduction, class balancing, and hyper parameter tuning to enhance CVD prediction accuracy. A combined dataset of 920 patient records from multiple clinical sources was cleaned, standardized, and processed using Principal Component Analysis (PCA) to reduce redundancy and multi collinearity among correlated medical attributes. Class imbalance was addressed using SMOTE, significantly improving minority class sensitivity. Six machine learning classifiers including Random Forest, Gradient Boosting, SVC, Decision Tree, AdaBoost, and an Ensemble Classifier were trained and evaluated under four configurations: with or without PCA and with or without hyper parameter tuning via Grid SearchCV. Results demonstrate that the integration of PCA and tuning consistently improvedmodel performance, with the Ensemble Classifier and Random Forest achieving accuracies and F1-scores. Significant improvements in recall and ROC-AUC values indicate better discriminatory power, especially for disease positive cases. Overall, the proposed pipeline provides a robust, generalizable approach for CVD prediction and highlights the importance of dimensionality reduction, data balancing, and parameter optimization in developing clinically reliable diagnostic models.

Keywords:Hyper-parameter tuning, Classification Algorithm, Prediction, Heart Disease, and Dimension reduction.

I. INTRODUCTION

The World Health Organization (WHO) estimates that 32% of fatalities worldwide are caused by cardiovascular diseases (CVDs) [1]. Early detection and prediction of cardiovascular risk have become more crucial in order to lessen the severe effects of CVDs. Conventional risk assessment models, like the SCORE system and the Framingham Heart Study, mostly rely on clinical and demographic variables including age, sex, blood pressure, and cholesterol. Despite their widespread usage in clinical practice, these models have some important drawbacks, such as their inability to identify intricate, non linear patterns in the data and their failure to take into account numerous unique health characteristics that could have a substantial impact on cardiovascular risk.

Machine learning (ML) techniques, with their ability to handle large and complex datasets, have gained significant attention in recent years as a potential solution for improving cardiovascular risk prediction. ML algorithms can automatically detect patterns in data that might not be obvious to human analysts, allowing for more accurate and personalized predictions. Researchers have been exploring various ML approaches, ranging from ensemble methods to deep learning, for the prediction of cardiovascular diseases, yielding promising results that outperform traditional models in many cases.

The research also covers studies that combine tree based methods with deep learning approaches, enabling the extraction of both linear and nonlinear patterns from data, which further improves predictive performance. Despite the significant advancements in these ML techniques, challenges remain, particularly with issues related to model interpretability, computational resources, and the need for high quality datasets. The integration of diverse health data, including psychological and sensor based inputs, presents a promising avenue for future research, aiming to overcome the current limitations of traditional and machine learning based models.

By synthesizing the findings from recent studies, this aims to provide an in depth understanding of the current state of cardiovascular disease prediction using machine learning techniques, offering insights into both the progress made and the

http://doi.org/10.63766/spujstmr.26.000071

challenges that lie ahead in this field.

II. EXISTING APPROACH

Because cardiovascular diseases (CVDs) remain one of the world's top causes of death, research into predictive modelling utilizing machine learning (ML) techniques is still underway. This looks at new research that highlights improvements in machine learning techniques for CVD prediction. As computational tools, machine learning methods (MLTs) are essential for the early diagnosis of cardiovascular illness. There are several methodical stages involved in estimating the risk of heart disease. Examining feature distributions, finding missing values, eliminating duplicate records, and fixing out of range entries are all part of the first stage, known as data preparation [30]. The collection is effectively managed and preprocessed using Pandas, a popular data analysis library [31].

Dimensionality reduction, data standardization, and hyperparameter optimization approaches are then used to improve model performance [27]. A prediction model is created to determine the probability of heart disease using the preprocessed and optimized data as input.Principal Component Analysis (PCA) and classification algorithms are fitted to the training dataset (X_train and Y_train) in a sequential manner. To find the best model configurations, hyperparameter tuning is done on the training data concurrently. Standard performance criteria, including as accuracy, precision, recall, and F1-score, are used to assess the efficacy of the suggested method.

Attributes	Definition and values
age	The person's age in years
sex	The person's sex (1 = male, 0 = female)
cp	The chest pain experienced (Value 1: typical angina, Value 2: atypical angina, Value 3: non-anginal pain, Value 4: asymptomatic)
trestbps	The person's resting blood pressure (mm Hg on admission to the hospital)
chol	The person's cholesterol measurement in mg/dl
fbs	The person's fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false)
restecg	Resting electrocardiographic measurement (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria)
thalach	The person's maximum heart rate achieved
exang	Exercise induced angina (1 = yes; 0 = no)
oldpeak	ST depression induced by exercise relative to rest ('ST' relates to positions on the ECG plot. See more here)
slope	the slope of the peak exercise ST segment (Value 1: upsloping, Value 2: flat, Value 3: downsloping)
ca	The number of major vessels (0–3)
thal	A blood disorder called thalassemia (3 = normal; 6 = fixed defect; 7 = reversable defect)
target	Heart disease (0 = no, 1 = yes)

III. IMPLEMENTATION STRATEGY

An organized and methodical machine learning pipeline is used to build the suggested cardiovascular disease (CVD) prediction system. To guarantee accurate and dependable predictions, the approach concentrates on data preparation, feature optimization, model training, hyperparameter adjustment, and performance assessment.

3.1 Methods of Classification

A supervised learning technique called data classification is used to give a given input instance the proper class label [32]. In order to facilitate effective data storage, retrieval, and analysis, a classifier groups the data into predetermined categories. Labelled training data is used to train the classification model, and test data that has not yet been seen is used to validate the model's performance.

3.1.1 RF

Using randomly chosen subsets of the training data, Random Forest is an ensemble based classification method that builds many decision trees. Every tree makes a forecast, and the average or majority vote of all the trees determines the ultimate result. Furthermore, RF prevents over fitting and improves prediction accuracy by selecting significant characteristics from the dataset [14].

3.1.2 GB

A machine learning method that is a member of the ensemble learning family is gradient boosting. It gradually combines several weak learners, usually decision trees, to create a strong prediction model. In order to increase classification performance, each new model is trained to fix the mistakes caused by the ones that came before it.

3.1.3 SVC

By locating an ideal hyper plane in a high dimensional feature space, the Support Vector Classifier seeks to divide data points into discrete classes. Kernel functions are used to alter the input data in order to do this. The Support Vector Machine (SVM) framework is used to implement SVC, which is well known for its efficacy in control and classification applications [16].

3.1.4 DT

A popular classification system called Decision Tree uses a hierarchical, tree like structure to represent decisions. The data is divided according to feature based conditions beginning at the root node and continuing until a leaf node is reached, which yields the final class label. For jobs involving decision making, this approach is simple to use and efficient [21].

3.1.5 AdaBoost

Adaptive Boosting, or AdaBoost for short, is an ensemble learning technique that improves classification performance by combining several weak learners, frequently straightforward decision stumps. By giving misclassified examples more weights, the algorithm improves overall accuracy by allowing later learners to concentrate on challenging data.

3.1.6 Classifier Ensemble

To generate a final prediction, an ensemble classifier combines several basic classifiers, which could be weak or strong learners. Ensemble learning is based on the core idea that a group of various models can outperform any single classifier in terms of accuracy and robustness, especially when individual models show distinct error patterns.

3.2 Techniques for Dimensionality Reduction

3.2.1 PCA

An unsupervised learning technique called dimensionality reduction converts high dimensional datasets into lower dimensional representations while preserving the key features of the original data. Improved model performance, decreased computational complexity, cheaper memory needs, and shorter processing times are just a few benefits of this technique [39].

A popular dimensionality reduction method that reduces information loss when converting a high dimensional dataset into a lower dimensional space is principal component analysis (PCA). There are four primary steps in the PCA method. To guarantee that every characteristic contributes equally to the study, the dataset is first normalized. The correlations between the variables are then captured by computing the covariance matrix of the standardized features.

The covariance matrix's eigen values and matching eigenvectors are then found. The top components are then chosen to create the reduced feature space after the eigenvectors are sorted according to the corresponding eigenvalues. PCA is an unsupervised learning technique that finds pathways that capture the most variation in the data without taking class labels into account [20, 21].

ISSN: 3049-1479(Online)

3.2.2 SMOTE

A data preparation technique called Synthetic Minority Over sampling Technique (SMOTE) was created to overcome class imbalance in machine learning datasets. Models that are trained on unbalanced data frequently develop bias in favor of the majority class, which impairs their ability to predict instances of the minority class. SMOTE creates new synthetic instances by interpolating between nearby minority class samples in the feature space, in contrast to conventional oversampling techniques that only replicate existing minority samples.

This evaluation's goal is to demonstrate how SMOTE can enhance machine learning models' accuracy and dependability. This is accomplished by evaluating a number of popular classifiers, such as Naïve Bayes (NB), Logistic Regression (LR), Rotation Forest (RotF), Multilayer Perceptron (MLP), K-Nearest Neighbors (KNN), J48, Bagging, Random Forest (RF), Voting, and Stacking. Ten-fold cross-validation is used to assess model performance using accuracy, recall, precision, and area under the ROC curve (AUC) metrics both with and without the use of SMOTE.

3.3 Hyperparameter Tuning

In order to optimize machine learning models for better prediction performance, hyperparameter tuning is an essential step. A popular method called Grid SearchCV methodically investigates a predetermined set of hyperparameter combinations to determine the best configuration for a particular model. The tuning technique in this study takes into account hyperparameters like the number of layers and the learning rate.

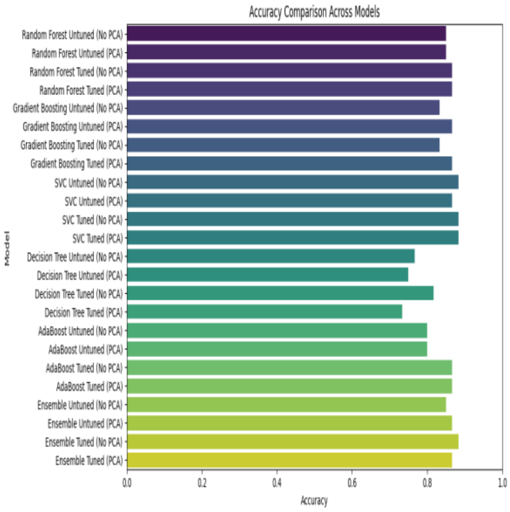
Grid SearchCV uses cross validation to assess every possible combination of the chosen hyperparameters and finds the configuration that produces the best results. The final model is then trained using the ideal collection of hyperparameters, guaranteeing improved accuracy and resilience [30].

IV. RESULT AND DISCUSSION

This study uses a combined dataset of 920 samples and 14 attributes from four popular cardiovascular databases: Long Beach, Virginia (200 samples), Cleveland (304 samples), Hungarian (293 samples), and Switzerland (123 samples) [20]. Thirty percent of the total samples are put aside as the test set for cardiovascular disease prediction in order to assess the prediction performance of the suggested models.

Table 1 summarizes the dataset attributes and their descriptions, and Figure 11 shows how each attribute is distributed in relation to the others. Two particular properties need to be corrected during preprocessing. First, the ca attribute had five initial values (0, 1, 2, 3, and 4); however, the value 4 was eliminated and replaced with the value 3 because it was out of range. Second, there are erroneous zero values in the slope attribute that are regarded as missing data. The slope attribute's median value is 1 after these missing values are imputed using the Simple Imputer technique using a median strategy.

The Standard Scaler method is used to normalize the feature set after the target variable has been divided into the input feature matrix (X) and output labels (y). This study uses six machine learning classifiers Random Forest (RF), Gradient Boosting (GB), Support Vector Classifier (SVC), Decision Tree (DT), AdaBoost (AB), and an Ensemble Classifier in conjunction with dimensionality reduction and optimization methods, such as Principal Component Analysis (PCA) and Grid SearchCV, to produce precise and trustworthy predictions.



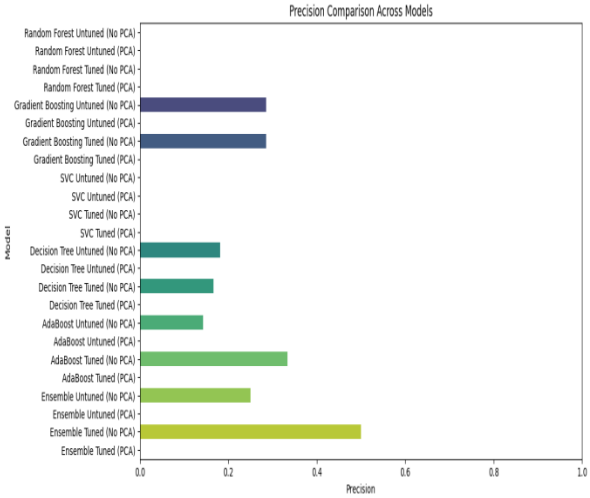
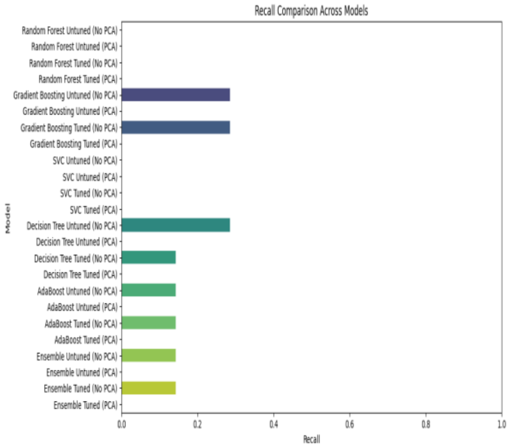


Fig. 1: Accuracy Comparison across Models

Fig. 2: Precision Comparison across Models



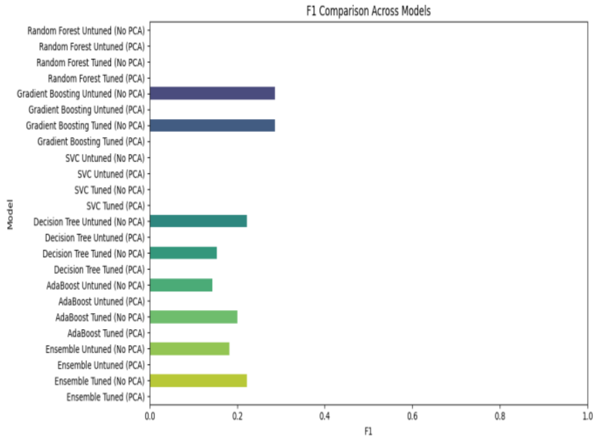


Fig. 3: Recall Comparison across Models

Fig. 4: F1-Score Comparison across Models

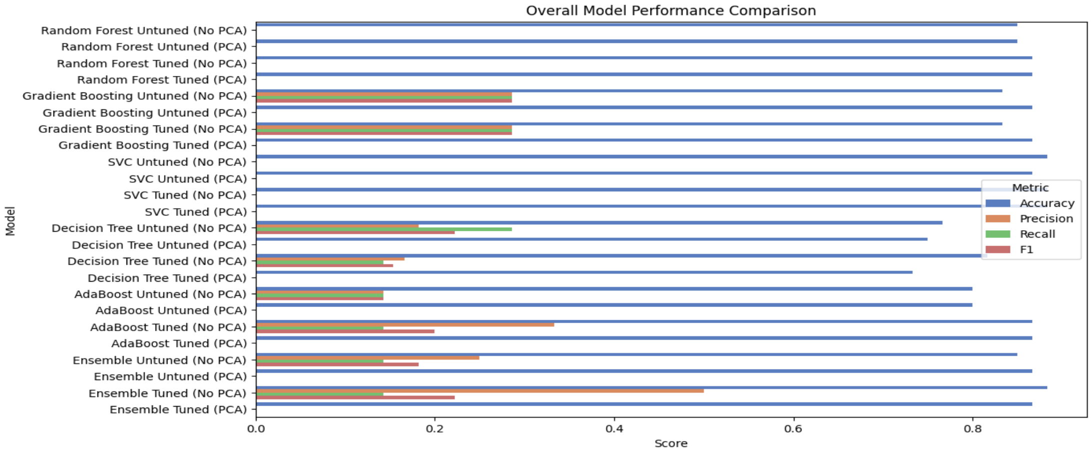


Fig. 5: Overall Models Performance Comparison

ISSN: 3049-1479(Online)

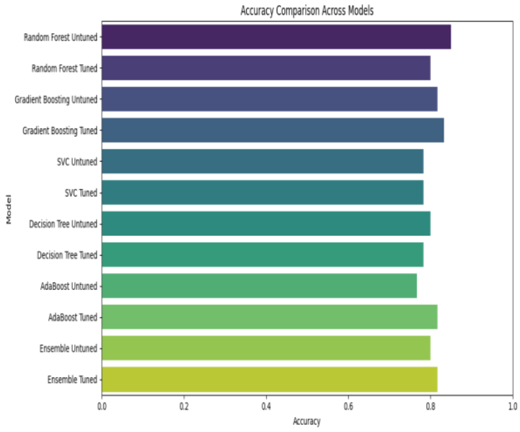


Fig. 6: Accuracy Comparison across Models

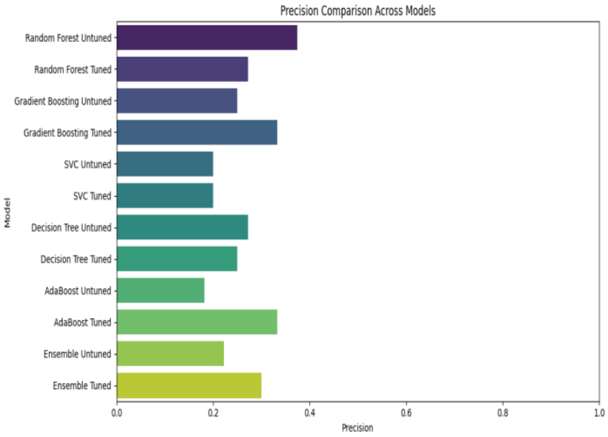


Fig. 7: Precision Comparison across Models

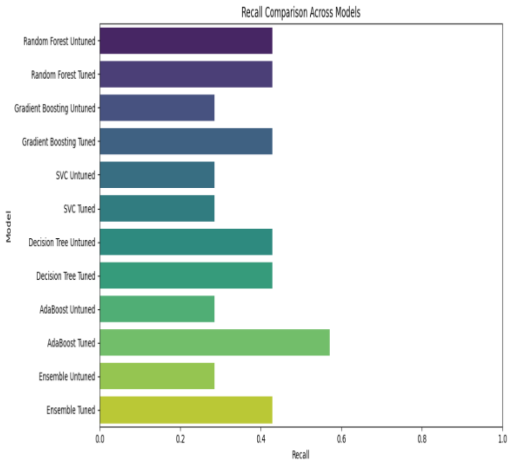


Fig. 8: Recall Comparison across Models

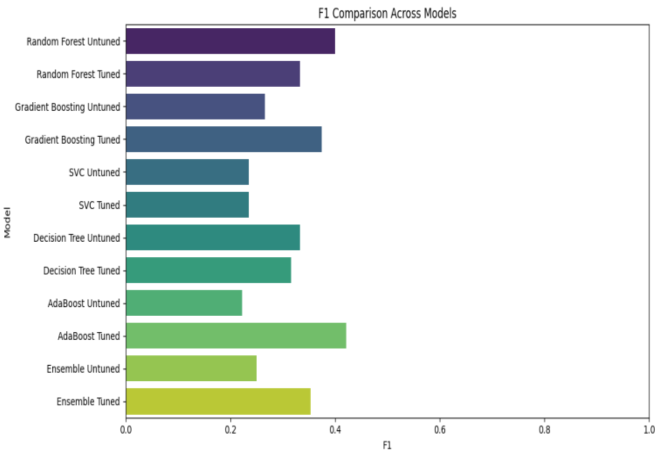
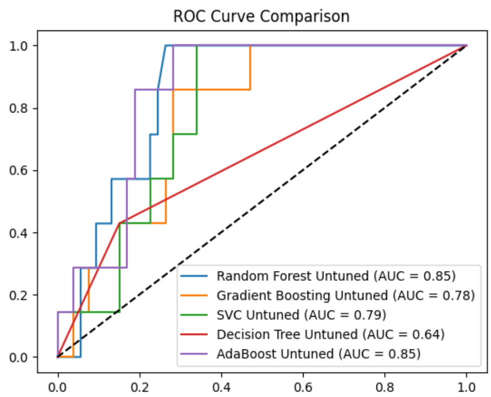
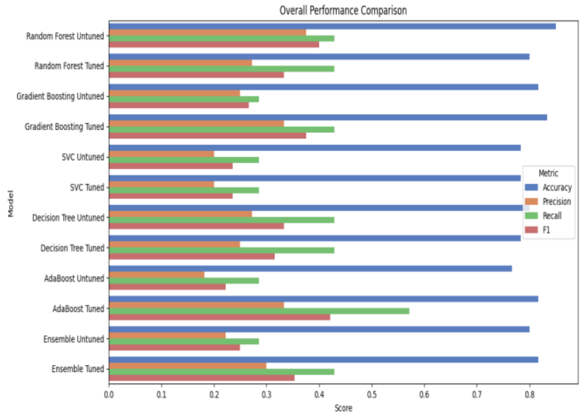


Fig. 9: F1-Score Comparison across Models





V. CONCLUSION

The integration of machine learning into cardiovascular disease prediction offers immense potential for improving early detection and personalized treatment plans. However, future research should focus on enhancing model interpretability, addressing data quality issues, and exploring the clinical feasibility of combining multiple data types, such as physiological and psychological factors, to provide more accurate and actionable predictions.

REFERENCES

[1] A. Tiwari, A. Chugh, and A. Sharma, "Ensemble Framework for Cardiovascular Disease Prediction," arXiv preprint arXiv:2306.09989, 2023. *[Online]. Available: https://arxiv.org/abs/2306.09989. [Accessed: Jan. 3, 2025].*

[2] M. T. García-Ordás, M. Bayón-Gutiérrez, C. Benavides, J. Aveleira-Mata, and J. A. Benítez-Andrades, "Heart disease risk prediction using deep learning techniques with feature augmentation," *arXiv preprint arXiv:2402.05495, 2024. [Online]. Available: https://arxiv.org/abs/2402.05495. [Accessed: Jan. 3, 2025].*

[3] W.-H. Weng, S. Baur, M. Daswani, C. Chen, L. Harrell, S. Kakarmath, M. Jabara, B. Behsaz, C. Y. McLean, Y. Matias, G. S. Corrado, S. Shetty, S. Prabhakara, Y. Liu, G. Danaei, and D. Ardila, "Predicting cardiovascular disease risk using photoplethysmography and deep learning," *arXiv preprint arXiv:2305.05648, 2023. [Online]. Available: https://arxiv.org/abs/2305.05648. [Accessed: Jan. 3, 2025].*

[4] M. Shishehbori and M. Awan, "Incorporating psychological data into machine learning models for cardiovascular disease prediction," *arXiv preprint arXiv:2401.17328, 2024. [Online]. Available: https://arxiv.org/abs/2401.17328. [Accessed: Jan. 3, 2025].*

[5] M. Hossain, M. Islam, A. Rahman, and S. Hoque, "A Comparative Study of Tree-Based Machine Learning Methods for Cardiovascular Disease Detection," *Scientific Reports, vol. 14, no. 1, p. 72819, 2024. [Online]. Available: https://www.nature.com/articles/s41598-024-72819-9. [Accessed: Jan. 3, 2025].*

[6] M. T. García-Ordás, M. Bayón-Gutiérrez, C. Benavides, J. Aveleira-Mata, and J. A. Benítez-Andrades, "Combining machine learning and deep learning techniques for cardiovascular disease prediction," Eur. J. Med. Res., vol. 29, no. 1, pp. 20-30, 2024. [Online]. Available: https://eurjmedres.biomedcentral.com/articles/10.1186/s40001-024-02044-7. [Accessed: Jan. 3, 2025].

[7] Organization, W.H. (2023) Cardiovascular diseases, World Health Organization. Available at: https://www.who.int/healthtopics/cardiovascular-diseases (Accessed: 23 May 2023).

[8] MadhukarChavan, Shashank Kumar Singh, SaurabhBansod and Prashant Pal, “Design and Implementation of Heart Disease Prediction Using Artificial Neural Network” 2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM) | 979-8-3503-4779-1/23/$31.00 ©2023 IEEE| DOI: 10.1109/ICONSTEM56934.2023.10142267.

[9] Barbara Martins, Diana Ferreira, Cristiana Neto, Antonio Abelha,Jose Machado , “Data Mining for Cardiovascular Disease Prediction,” ©Springer, 5 January 2021 .

[10] S. Patel, J. Patel and S. Tejalupadhyay,"Cardiovascular Disease prediction using Machine learning and Data Mining Technique", *Journal - IJCSC, vol. 7, 2022. Available: 10.090592/IJCSC.2016.018.*

[11] Malav, A., Kadam, K. and Kamat, P. prediction of heart disease using K-means and artificial neural network Available*at:https://www.researchgate.net/publication/319486202_Prediction_Of_Heart_Disease_Using_KMeans_and _Artificial_Neural_Network_as_H Hybrid_Approach_to_Improve_Accuracy (Accessed: 22 May 2023).*

[12] NajmulHasan and YukunBao, “Comparing different feature selection algorithms for cardiovascular disease prediction*,” © Springer,October 2020.*

[13] M. A. Alim, S. Habib, Y. Farooq and A. Rafay 2020 Robust Heart Disease Prediction: A Novel Approach based on Significant Feature and Ensemble learning Model 2020 *3rd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), pp 1–5.*

[14] Dr. Joy IongZong Chen, P. Hengjinda, “Early Prediction of Coronary Artery Disease (CAD) by Machine Learning Method-A Comparative Study” *Journal of Artificial Intelligence and Capsule Networks (2021) Vol.03/ No.01 http://irojournals.com/aicn/DOI: https://doi.org/10.36548/jaicn.2021.1.002.*

[15] DivyaAnnepu, Gowtham G, "Cardio vascular Disease Prediction Using Machine Learning Techniques" *International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 06 Issue: 04 | Apr 2019 www.irjet.net p-ISSN: 2395-0072.*

[16] KaziMostaqhridoy, Saiful Islam Akash, FahmidaAfroseDipti, MahamudulHasan, JesanAhammedovi, Md Al-Imran, Dr. SarwarJahan "Heart Disease Algorithm using Machine Learning Algorithms" *2023 4th International Conference on Big Data Analytics and Practices (IBDAP) | 979-8-3503-0019- 2/23/$31.00 ©2023 IEEE | DOI: 10.1109/IBDAP58581.2023.10271997*

[17] S. Ismaeel, A. Miri and D. Chourishi 2015 Using the Extreme Learning Machine (ELM) technique for heart disease diagnosis *2015 IEEE Canada International Humanitarian Technology Conference (IHTC2015), pp 1–3*

[18] Abdullah Alqahtani, ShtwaiAlsubai, Mohammed Sha, Lucia Vilcekova, and TalhaJaved, "*Cardiovascular Disease Detection using Ensemble Learning," Hindawi, August 2022.*

[19] Dangare, C. and Apte, S. (2012) A data mining approach for prediction of heart disease using Neural Networks, SSRN. *Available at: https://papers.ssrn.com/sol3/papers.cfm?abstractid=2175569 (Accessed: 22 May 2023).*

[20] Welcome to the UC Irvine Machine Learning Repository (no date) UCI Machine Learning Repository. *Available at: https://archive.ics.uci.edu/ (Accessed: 19 June 2023).*

[21] https://www.kaggle.com/sulianova/cardiovascular-disease-dataset

[22] C. J. Harrison and C. J. S. Gibbons, "Machine learning in medicine: a practical introduction to natural language processing," *BMC Medical Research Methodology, vol. 21, no. 1, p. 158, 2021.*

[23] R. Kumar, P. Kumar, R. Tripathi, G. P. Gupta, A. K. M. N. Islam, and M. Shorfuzzaman, "Permissioned blockchainanddeep- *learning for secure and efficient data sharing in industrial health care systems," IEEE Transactions on Industrial Informatics, p. 1, 2022.*

[24] Karl PatrikKresoja, Matthias Unterhuber, Rolf Wachter,HolgerThiele,PhilippLurz, "A cardiologists guide to machine learning in cardiovascular disease prognosis prediction" *Basic Research in Cardiology (2023) 118:10 https://doi.org/10.1007/s00395-023-00982-7 ©Springer, 20 March 2023.*

[25] Sibo Prasad Patro, NeelamadhabPadhy, DukuruChiranjevi, "Ambient assisted living predictive model for cardiovascular disease prediction using supervised learning" *Evolutionary Intelligence https://doi.org/10.1007/s12065-020-00484-8, © Springer-Verlag GmbH Germany, part of Springer Nature 2020.*

[26] SaravananSrinivasan, SubathraGunasekaran, SandeepKumarMathivanan, BenjulaAnbu Malar M. B, PrabhuJayagopal&GemmachisTeshiteDalu, "An active learning machine technique-based prediction of cardiovascular heart disease from UCI repository database"Scientific Reports | (2023) 13:13588 | https://doi.org/10.1038/s41598-023-40717-1.

[27] MadhukarChavan, Shashank Kumar Singh, SaurabhBansod and Prashant Pal, "Design and Implementation of Heart Disease Prediction Using Artificial Neural Network" *2023 Eighth International Conference on Science Technology Engineering and Mathematics(ICONSTEM) | 979-8-3503-4779- 1/23/$31.00 ©2023 IEEE| DOI: 10.1109/ICONSTEM56934.2023.10142.*

[28] GhulabNabiAhamad, Shafiullah ,Hira Fatima, Imdadullah ,S. M. Zakariya, Mohamed Abbas,Mohammed S. Alqahtani ,and Mohammed Usman "*Influence of Optimal Hyperparameters on the Performance of Machine Learning Algorithms for Predicting Heart Disease" Processes 2023, 11, 734. https://doi.org/10.3390/pr11030734.*

[29] GamalSaad Mohamed Khamis and Sultan MunadiAlanazi " Exploring sex disparities in cardiovascular disease risk factors using principal component analysis and latent class analysis techniques" *Khamis and Alanazi BMC Medical Informatics and Decision Making (2023) 23:101 https://doi.org/10.1186/s12911-023-02179-3.*

[30] Karna Vishnu Vardhana Reddy, IrraivanElamvazuthi, AzrinaAbd Aziz, SivajothiParamasivam, Hui Na Chua and SatyamurthyPranavanand "*An Efficient Prediction System for Coronary Heart Disease Risk Using Selected Principal Components and Hyperparameter Optimization" Appl. Sci. 2023, 13, 118. https://doi.org/10.3390/app13010118.*

[31] Maria Trigka and Elias Dritsas "Long-Term Coronary Artery Disease Risk Prediction with Machine Learning Models*" Sensors 2023, 23,1193. https://doi.org/10.3390/s23031193.*

[32] Abdullah Alqahtani ,ShtwaiAlsubai , MohemmedSha , Lucia Vilcekova, and TalhaJaved , "Cardiovascular Disease Detection using Ensemble Learning, " *Hindawi Computational Intelligence and Neuroscience Volume 2022, Article ID 5267498, 9 pages https://doi.org/10.1155/2022/5267498,August 2022.*