Heart Disease Prediction with Feature Engineering, SMOTE Augmentation, and Interpretable Deep Learning Models

Abdul Hamid
M.Tech Scholar
Department of Computer Science and Engineering
Sam Global University, Raisen
Bhopal, M.P, India
abdulhamid70086@gmail.com

Dr.Saurabh Mandloi
Head of Department
Department of Computer Science and Technology
Sam Global University, Raisen
Bhopal,M.P, India

Abstract: Cardiovascular disease is still a leading destination in all of the universe, and that's why this serious matter demands the development of effective and accurate forecasting methods to address the issue both at early stage and prevent it in the case of concern. This work utilizes the machine learning and deep learning paradigms to manipulate the Cleveland Heart Disease dataset (CHDD) in order to make easy and credible predictions of heart disease. Working with the data included pre-processing such as missing value replacements, outlier detection and removal, standardisation, as such – the authors balanced the classes using SMOTE (Synthetic Minority Over-sampling Technique). Apart from that, various techniques such as Analysis of variance (ANOVA), Chi-square, and Mutual Information were employed for feature selection lead to improved subsets (SF-1, SF-2, SF-3) which were classifiers subjected to tests including Logistic Regression, Support Vector Machines, K-Nearest Neighbours and Voting Ensembles, among others. The outcomes revealed that the unsophisticated together with single classifiers were able to obtain around 90-91% accuracies. On the other hand, the cutting-edge Regularized Deep Feedforward Neural Network (DNN) with Swish activation, AdamW optimizer and SMOTE oversampling boosted the accuracy rate substantially, registering an impressive accuracy figure of 98% with balanced precision, recall and F1-score. SHAP explainer was used to improve the model interpretability and the final model was packaged as a mobile application for medical professionals to use in real time. The understanding is that in contradiction to individual ML or DL models, there is potential of hybrid ML-DL pipelines in implementing applications, which use cardiovascular risk prediction and are reliable, affordable, and can be easily scaled up.

Keywords: Heart Disease Prediction, Machine Learning, Deep Learning, Cleveland Heart Disease Dataset (CHDD), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Voting Ensemble, Feature Selection, SMOTE, Regularized Neural Network, SHAP, Mobile Health Application.

I. INTRODUCTION

Heart conditions are some of the foremost causes of mortality worldwide, greatly straining societies and healthcare systems [1]. It comprises ailments like coronary artery disease, heart failure, and arrhythmias. With the rising trends of urbanization, sedentary lifestyles, and aging population, the prevalence of heart diseases keeps increasing with no end in sight, requiring foretelling and preventive measures [2].Figure 1.describes heart diseases.

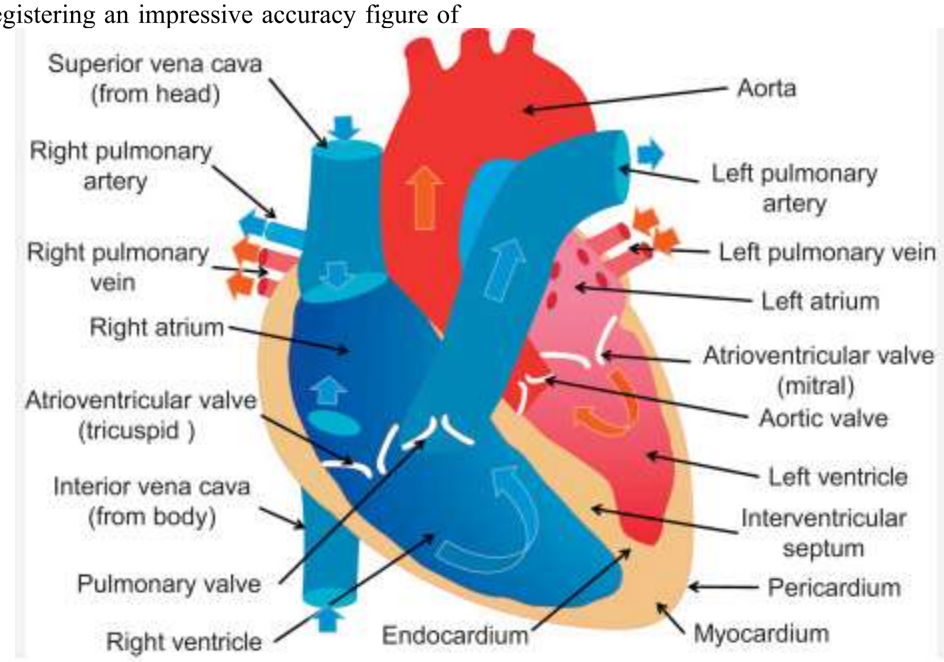

Figure 1. Heart Diseases

Cardiovascular diseases (CVDs) continue to be the number one death world in the whole world, accounting for almost one-third of all deaths with seventeen point nine million people dying yearly following the heaviest toll felt on countries with low and middle-level incomes. There is seen a rise in the number of these diseases and this is where it is due to urbanization, poor dietary habits and low awareness among other factors and as such the cost of treatment and productivity foreigners are also a point of contention [4]. Risks of acquiring this disease can also be due to some factors that can be modified and cannot be modified such as. Also, unhealthy practices like unhealthy diet or smoking and excessive drinking and also Alcoholism itself, stress, increased body weight, hypertension as a disease, diabetes and poor quality of sleep are very common. In terms of treatments, it could take a lifetime to educate patients and fine tune their habits compared to curative measures. Prevention is about complex activities such as reducing the behaviour that increases risk of disease, balanced intake of foods, and cancer screening at an appropriate age, as in the case of prevention of other chronic diseases, so these methods work [5][6]. Due to the fact of formal recognition of disease-associated related factors overtime, risking having complications such as myocardial infarction, cerebrovascular incidents, and sudden cardiac death, therefore, the process of early detection is of monumental concern and machine learning-based computational models hold promise for analysing and processing large volumes of data collaboration to detect patterns that are not visible and, at the same time, minimize the cost of medical care within reach [7][8].

A. Challenges in Traditional Diagnosis

The progressive approach to duration of resources on heart-related diseases encounters insuperable problems. These may range from reliance on the expertise of health specialists and the accuracy of most primary diagnostic tools to slow identification of signs presented in the early stages of these diseases [9][10]. Nonetheless, even these essential clinical judgments may contain contradictions due to some common socio-psychological features, or even the lack of professionally trained cardiologists in poor-complex organizations (countries), therefore artificial intelligence is required in the form of models of machine-learning to guide and support decision-making modes further [11][12]. Classical instruments such as ECG, ultrasound, angiography or blood tests often notice the presence of a disease only after there are obvious clinical symptoms. And there are associated risks with the use of the intervention of these methods for the management of the disease where it is not suitable and they are not possible—early signs are not captured [13][14]. This delay in detection heightens the risk of severe impacts such as heart attacks resulting in myocardial infarction and sudden death which frustrates the purpose of treatment. Healthcare resources no longer available appreciably constrain health care costs and lower patient outcomes stressing the critical need for data-enhanced recall and reduction on death as opposed to visit care [15][16].

B. Role of Data Science in Healthcare

Healthcare has witnessed an immense transformation over the last few years with the budding discipline of advanced health analytics solutions to the ever-burgeoning healthcare space. This is achieved through the creation of statistical models among other machine learning and predictive tools with the ability to consume large clinical records to enhance diagnostics and guide treatment. This therefore cuts across medical principles and practice to software practice [17]. With the onset of huge volumes of electronic data largely-driven by electronic health records, imagery, medical devices, genomic mapping and even patient- specific encounter logs has revolutionized the understanding of patient backgrounds, disease trajectories and also how certain types of interventions work in the era of real time forecasting, large scale epidemiologic studies, and individualized patient care [18][19]. When combined with the modern analytical tools and cloud based platforms, the health systems research would be observing unparalleled advancement, where efficient operational costs and research would take considerable steps. Similarly, A.I. and D.S.S systems provide components for the treatment of evidence-based clinical practice, doctor-patient intercommunication, healthcare products, and cost control, which has the power of removing problems, standardizing treatments, and improving multitasking [21]. DSS uses predictive analytics to advance care not only curative but preventive value enhancement, cost containment, and optimal use of healthcare utilization resources and hence becoming critical in improving patients' conditions and moving delivery of health services to current levels [22][23].

C. Machine Learning for Heart Disease Prediction

Machine learning undoubtedly contributes to the prediction and prevention of heart diseases. In particular, machine learning uses patient data, lifestyles with an emphasis on certain habits, and clinical pictures to predict people who are at risk and to prevent diagnosis mistakes and apply individual therapies in time [24]. In particular, the employment of classification models aimed at supervised learning approaches, such as such as logistic regression, support vector machines (SVM), decision trees, and random forests helps in sorting patients by the degree of risk with the use of such risk factors as cholesterol, blood pressure, age, and others, which is conducive for the improvement of the quality of preventive measures for cardiovascular diseases and treatment [25][26]. Furthermore, many times the impairment of the ability to standardize the classification model in an effort to classify the patient without the proper risk labels is experienced. This why another group of methods has been refined comprising techniques that operate in detachments (e.g. those that will not demand risk information) and k-means clustering and semi-supervised grouping were purposed which utilizes the unlabeled data history and case records of patients in such a way that the patterns are difficulty hidden principles concerning patient risk and therapeutic sub-types otherwise secretive case of diseases [27][28]. Correspondingly, a relevant and very crucial aspect is that of the selection of the most suitable features. For instance, it is possible to employ correlation analysis, PCA, recursive feature elimination, and other feature selection techniques in order to get rid of extraneous or irrelevant attributes, accentuate key risk factors, and most significantly elevate the performance and appropriateness of the predictions as well as make the modeling more convenient, by means of focusing on certain medical features [29][30].

D. Research objective

- Crafting and deploying predictive machine learning models that assess the heart disease risk using biomedical and demographic data.
- Evaluating and contrasting the performance of different classification algorithms such as Logistic Regression, Random Forest, SVM, and Neural Networks, on the basis of accuracy, precision, recall, or F1-score.
- Applying feature selection and evaluation methods on the data to identify the key risk factors for heart disease.

II. LITERATURE REVIEW

Recently conducted research has broached numerous classifications of machine learning for heart disease prediction such as the random forest (RF) and relative comparisons. A study by **M. G. El-Shafiey et al. [1] (2022)** detailed development of GAPSO-RF, a feature selector that combined GA and PSO with RF and proved high precision levels of 95.6 and recall of 91.4%, but such precision came at the cost of computation intensity and inadequate validation. **Goswami et al. [2] (2022)** similarly introduced a cost-sensitive RF variant in anticipation of common medical imagery training imbalances, where the cost was guided by clinicians positions, while **Heidari et al. [3] (2022)** took the evaluation of RFC further by using alternate route and a quantum neural network pipeline, where the partitioned RF was shown to be more scalable than the quantum neural network; **Friis [4] (2015)** enriched the analysis by also considering the basic machine learning EURODIVITY and soft-voting coherent portfolios with random forest bases, the author shows that when either is used with otherwise standard RF models, the risk of overfitting may be high. Who recommended RF-based soft-voting ensembles, and **Suresh [5] (2021),** who applied swarm optimization to RF tuning, though both risked over fitting and complexity. **Chang [6] (2022)** implemented kernelized random forests with black hole optimization to have improved sensitivity with reduced interpretability, while **Shahrokh Asadi et al. [7] (2021)** proposed that the optimization must be carried out on the random forest without introducing new methods. Other studies compared DTs when balanced datasets were available **(S. A. Alluhaidan et al. [8] (2022)),** the accuracy of hands-on weighed RF on a rather rough dataset of Cleveland heart disease's **(D. A. Hossen et al. [9] (2021)),** or developed an optimization and boosting for RF **(Madhumita Pal et al. [10] (2022)).** Apart from RF, a few studies analyzed logistic regression (LR) and naïve bayes (NB): **Kwakye and Dadzie [11] (2021)** reported that NLP data proved to be more robust and that the other methods were always less efficient while NB was good but less predictive on all UCI data; **Subramanian et al. [12] (2022)** provided further confirmation to remember logistics regression was accurate was used on the Indian local standard; **Surya et al. [13] (2021)** stated that RF+SVM did slightly better in terms of ensembles compared to LR/NB alone; **Kumar et al. [14] (2021)** found that LR could recover approximately 75% accuracy and NB around 70%. In the studies of LR vs. NB it was shown that LR was marginally better and more interpreta- ble than NB across the datas- ets **(M. A. Javeed et al. [15] (2021))** whereas NB outperformed LR in pre-natal MRI based detection **(Naray- ana & Nalini [16] (2021))** and structured data- sets **(Saxena et al. [17] (2021)).** Nearly 88% accuracy, with respect to smart health prediction systems supported by NB, has also been reported **(Saraswat et al. [18] (2023)),** while evaluations have emphasised LR's interpretability **(Ahsan & Siddique [19] (2021)**; **Sahoo et al. [20] (2021)**) and NB's computational efficiency, although almost always hybrids came out ahead of the Author Teams.

Table 1 Comparative Analysis of Machine Learning Approaches for Heart Disease Prediction

Author(s) & Year	Method(s) Used	Findings	Limitations
M. G. El-Shafiey et al. [1] (2022)	GAPSO (Genetic + PSO) + RF	Hybrid wrapper improved accuracy (95.6% / 91.4%) on UCI datasets.	Heavy computation; limited external validation.
Goswami et al. [2] (2022)	Cost-sensitive RF + feature elimination	RF variant handled imbalanced datasets; reduced misclassification.	Relies on small datasets; cost setting needs clinician input.
Heidari et al. [3] (2022)	Partitioned RF + quantum-hybrid pipeline	Partitioning improved runtime; RF remained competitive.	Quantum-hybrid parts experimental; results dataset-limited.
Suryathe Aditya [4] (2021)	LR, SVM, DT, RF, MLP + soft-voting ensemble	Ensemble (RF-heavy) improved stability vs single DT; preprocessing/oversampling explored.	Limited peer review; small academic datasets only.
Suresh [5] (2021)	Swarm optimization + RF	Metaheuristic tuning improved performance over untuned RF and DT.	Risk of overfitting; higher runtime/complexity.
Chang [6] (2022)	Kernelized RF + black-hole optimization	Higher sensitivity; feature-importance explored.	Complex implementation; reduced interpretability.
Shahrokh Asadi et al. [7] (2021)	RF vs DT and others	RF robust on tabular clinical features; emphasized validation.	No novelty; relied on retrospective/public datasets.

S. A. Alluhaidan et al. [8] (2022)	DT + oversampling	DT with resampling competitive, interpretable for clinicians.	Oversampling risk of leakage; limited datasets.
D. A. Hossen et al. [9] (2021)	RF vs DT, SVM	RF achieved ~86–89% accuracy on Cleveland dataset with ROC/feature analysis.	Single dataset focus; lacks multi-cohort validation.
Madhumita Pal et al. [10] (2022)	Optimized RFs (GA/PSO/ensembles)	RF + feature selection/tuning > DT, non-tuned models; interpretable.	Heavy tuning; mostly UCI datasets; little external validation.
Kwakye & Dadzie [11] (2021)	LR, NB, others	LR outperformed others on original UCI data; NB efficient but less accurate.	Limited to UCI datasets, not real-world.
Subramanian et al. [12] (2022)	LR, NB, others	LR more reliable; NB faster but less accurate on local Indian dataset.	Small, local dataset; not generalizable.
Surya et al. [13] (2021)	LR, NB, Ensembles (RF+SVM)	Ensembles > base learners; LR & NB strong but weaker alone.	Small benchmark datasets only.
Kumar et al. [14] (2021)	LR, NB, KNN, RF	LR ~75% accuracy; NB ~70%; RF higher accuracy.	LR/NB less accurate; dataset limitations.
M. A. Javeed et al. [15] (2021)	LR vs NB on Cleveland, Statlog, Framingham	LR 77–91%, NB 75–88%; LR slightly better & interpretable.	Accuracy dataset-dependent; generalizability uncertain.
Narayana & Nalini [16] (2021)	LR & NB on MRI images	NB 88% > LR 84% for prenatal detection.	Small imaging dataset; limited generalizability.
Saxena et al. [17] (2021)	LR, NB, DT, RF, KNN	Gaussian NB > LR on Cleveland dataset.	Project-level study; no clinical depth.
Saraswat et al. [18] (2023)	Naïve Bayes	Built smart prediction tool; ~88% accuracy.	Software-oriented, not clinically validated.
Ahsan & Siddique [19] (2021)	Literature review of ML methods	LR most used (interpretability); NB efficient but less accurate.	No new model; only synthesized results.
P. K. Sahoo et al. [20] (2021)	Probabilistic classifiers (NB focus)	NB competitive on small datasets; LR as baseline.	No direct LR vs NB focus; limited comparisons.

III. RESEARCH METHODOLOGY

The given analysis assumed the utilization of the recalibrated form of the Cleveland Heart Disease Dataset (190 patient records and 54 clinical indicators). These encompass factors of, demographics – age, sex; physical, i.e. blood pressure and liver tests; examinations – fasting glucose, chest pain, ECG, and maximal heart rate; activity indicators – exercise-induced angina, ST depression, ST elevation, and slope; anthropometric measurements – white blood count, Red blood cell count, LDL-cholesterol, serum cholesterol in mg/ decilitre in serum, and the back and sides of left ventricle – fluoroscopic vessel angiography, thalassemia. The target variable was considered as binary indicating presence of disease or absence with 1 and 0, respectively. The source dataset was incomplete, consequently, the empty cells were dropped, and the data was reorganized in a consistent formatting and the processing focused on data management for ease of processing, predictive modeling or data analysis.

A. Implementation Tools

The heart disease prediction system was created using the Scikit-learn library from Python to write algorithms and select certain features. The Pickle and NumPy in model record and data manipulation respectively Also, Pandas for data cleaning, Matplotlib and Seaborn for visualization. Furthermore, deep learning has been fused with machine learning in a mobile e-health solution, which is aimed at facilitating preventive healthcare in an on-demand setting. Major classification models in this project are support vector machines (SVM) and k-nearest neighbors (KNN) with feature Sf-1, Sf-2 and Sf-3 combinations. SVM (Model A2) Multiple kernels were used: polynomial, RBF, and sigmoid, and hyperparameters tuning (C, γ, kernel) but, the model was found to be very sensitive to noise. For KNN, each point in a dataset votes to its own class using Euclidean or Minowski distance. Ensemble learning was investigated to reduce the bias, while normalization with SMOTE helped tackle the imbalance in the classes. Functional subunits are integrated in the deep learning pipe-lines for data cleaning, bias reduction, and classification algorithms, resulting in a sophisticated and efficient heart disease prediction system.

B. Data Pre-processing

The data pre-processing pipelines began by importing the health information of heart disease patients into the Pandas Data Frame and cleansing them by removing garbage values and incomplete files, as well as turning the target classification into binary class, i.e. 1 (presence of disease), 0 (absence). All other attributes were normalized using the Robust Scalar to cope with extreme values which are peculiar to clinical datasets thus attempting to make the learning and estimation faster and more robust. Finally, domain expert features were designed for example an interaction feature between the occurrence of exercise-induced angina and the change in ST elevation, as well as product of blood pressure and heart rate, vessel score and the age-standardized cholesterol index were built for the dataset. Such targeted feature manipulation helps in providing more meaningful details into the model to assist in learning.

C. Model Development

Regularized Deep Feedforward Neural Network and Swish activation function was used in the prediction model. This was backed up by AdamW optimizer and Regularization (L1, L2, dropout) To improve the generalization. It should be noted that Batch Normalization was also employed in order to stabilize training, ReLU was used for capturing non-linear patterns while sigmoid was chosen for binary classification. For data pre-processing, cleaning, feature engineering, SMOTE to overcome class imbalance, and scaling with Quantile Transformer addressing outliers were carried out. Other techniques that vintage class weighting, early stopping and gradient rescale achieved more efficient optimization and less over fitting.

The model assessing how correct and predictive it is was mainly based on accuracy, precision, recall, F1-score, AUC-ROC, and confusion matrices. These are the main parameters used to measure model efficiency.

Table 2 Description of Feature Set

S.No.	Feature Name	Description
1	Age	Patient's age in years
2	Sex	Gender of the patient (1 = male, 0 = female)
3	Chest Pain Type (CP)	Type of chest pain experienced (e.g., typical angina, atypical, non-anginal)
4	Resting Blood Pressure (trestbps)	Resting blood pressure in mm Hg
5	Serum Cholesterol (chol)	Serum cholesterol level in mg/dl
6	Fasting Blood Sugar (fbs)	Fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
7	Resting ECG Results (restecg)	Results of resting electrocardiograph (normal, abnormal, etc.)
8	Maximum Heart Rate Achieved (thalach)	Max heart rate achieved during exercise
9	Exercise-Induced Angina (exang)	Indicates exercise-induced angina (1 = yes; 0 = no)
10	Oldpeak	ST depression induced by exercise relative to rest
11	Slope of Peak Exercise ST Segment (slope)	The slope of the peak exercise ST segment (upsloping, flat, downsloping)
12	Number of Major Vessels Colored (ca)	Number of major vessels (0–3) colored by fluoroscopy
13	Thalassemia (thal)	Type of thalassemia (normal, fixed defect, reversible defect)

IV. RESULT AND DISCUSSION

This section gives a thorough performance evaluation of the proposed heart disease prediction framework across several ML classifiers in terms of efficiency, accuracy, and reliability of predictions. The experiments were carried out on the Cleveland Heart Disease Dataset, pre-processed for data consistency, and were further subjected to SMOTE wherein the class imbalance was addressed. Model evaluation was further executed with optimized feature subsets (SF-1, SF-2, and SF-3), allowing an evaluation of classifiers' predictive abilities under the various feature selection scenarios.

Accuracy is simply the proportion of TP and TN over the total samples, and its percentage may not always prove to be accurate in cases where the dataset is imbalanced. It is therefore necessary to use a proper ratio of other parameters alongside accuracy for a more perfect performance evaluation.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Sensitivity (Recall) refers to the model's capability to correctly identify patients with heart disease, i.e., the proportion of true positives detected among all actual positive cases.

$$Sensitivity = \frac{TP}{TP+FN} \quad (2)$$

Specificity is the ability of the model to correctly classify people without heart disease (true-negatives), in order to minimize false-positives and guarantee efficacy, cost efficiency, and trust from patients while keeping the need of sensitivity in check.

$$Specificity = \frac{TN}{TN+FP} \quad (3)$$

Precision is the proportion of true positives to predicted positives while ensuring that heart disease predictions are correct by including only positive test results and without causing harm, panic or frustration due to the excessive treatments.

$$Precision = \frac{TP}{TP+Fp} \quad (4)$$

The F1 Score is the harmonic mean of precision and recall, offering a fair measure for imbalanced data by balancing false positives and false negatives.

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (5)$$

AUC measures a model's ability to distinguish positives from negatives across thresholds, with higher values indicating stronger performance, especially useful for imbalanced datasets and critical in medical diagnostics.

$$TRP = \frac{TP}{TP+FN} \quad (6)$$

$$FPR = \frac{FP}{FP+TN} \quad (7)$$

A. Model Accuracy

In evaluating machine learning classifiers for heart disease, metrics like Accuracy, Precision, Recall, and F1-score and FPR were available. The Voting Classifier and SVM each achieved 90% accuracy. However, SVM saw an increase in Recall but balanced the Precision with KNN having slightly formed an impressive 91% accuracy with a remarkable Recall (0.93), therefore enabling easy identification. The Hybrid Voting Model, on the other hand, displayed consistent performance across performance metrics, with little to no over performance.

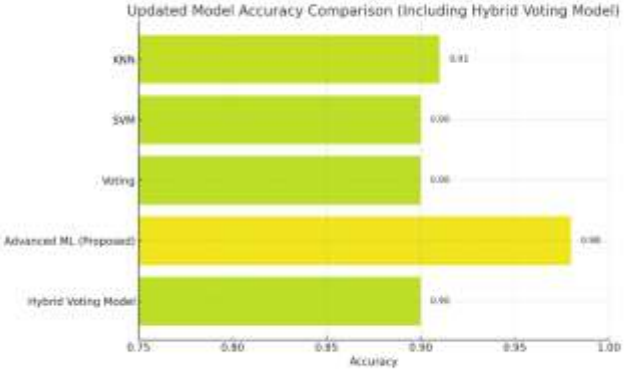

Figure 2 Model Accuracy Comparison: Baseline vs. Advanced Deep Learning

In terms of which model fared the best, it was the more advanced version of the Deep Neural Network (DNN) featuring feature engineering as well as SMOTE, that scored a massive 98% accuracy rate and all rates of precision, recall, and F1 score at 0.98 with the lowest FPR. Swish activation, AdamW optimization, and augmentation forced convergence and alleviated the model errors, which further illustrated the efficiency and reliability of the DNNs as opposed to the usual models.

A study was performed using a split of 75% for training and 25% for testing to avoid biased assessments. Through the 'train-test' validation are the ten machine learning methods combined with various frame conditions to find the best method for risk prediction of heart disease. The survey made evident operation and lack of it for each model, justified by the relevancy of features that have been used to setup different subspaces and choose the model. This enabled to clearly specify the usefulness of the algorithms and the respective feature vectors for better forecasting as shown in Figure 3.

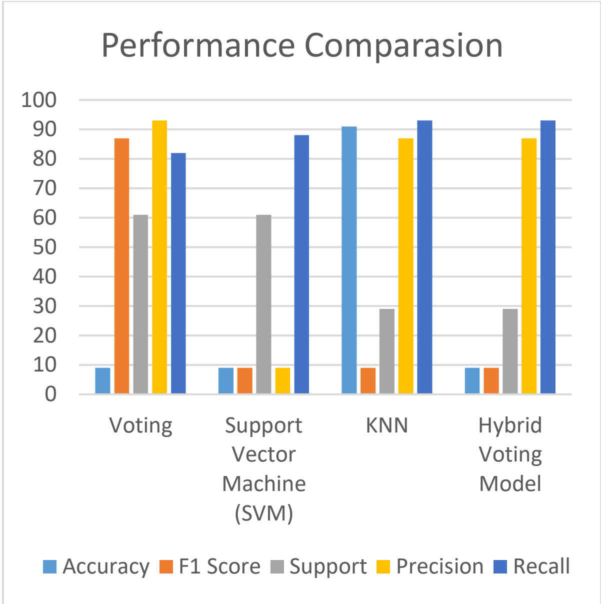

Figure 1 Performance Comparisons

Table 3 Performance Metrics of Regularized Deep Feed forward Neural Network with Swish Activation using AdamW Optimizer

Metric	Score
Accuracy	0.98
Sensitivity	96.61
Precision	0.97
F1 Score	0.97

V. CONCLUSION

The analysis supports the concept of a careful combination of conventional methods and advanced deep learning systems in terms of predicting heart disease. Although quite satisfactory between 90% and 91%, the baseline classifiers such as SVM, KNN and ensemble models promise very good results. At the same time, Regularized Deep Feedforward Neural Network with Swish activation and AdamW optimizer surpassed previously available classifiers, offering the accuracy of around 98% with AUC, sensitivity and IMP equal to 1 and matching well against class imbalance. Performance, clinical relevance and increasing transparency were enhanced by several device; feature engineering, SMOTE augmentation, SHAP interpretability. In addition, the integration of this system into a mobile tool demonstrates it is readily usable within real-time diagnosis for the benefit of doctors and patients providing them with the ability to immediately obtain an accurate and understandable assessment of cardiovascular risk. It is also necessary to note that further study needs to address the applicability of the presented model design on relatively wider and more different clinical datasets in order to ensure wider applicability and scalability.

Conflict of Interest: The corresponding author, on behalf of second author, confirms that there are no conflicts of interest to disclose.

References

[1] M. G. El-Shafiey, A. Hagag, E. S. A. El-Dahshan and M. A. Ismail, "A hybrid GA and PSO optimized approach for heart-disease prediction based on random forest," *Multimedia Tools and Applications*, vol. 81, no. 13, pp. 18155–18179, Mar. 2022. doi:10.1007/s11042-022-12425-x. ACM Digital LibraryResearchGate

[2] *Application of Double Sensitive Cost Random Forest in Heart Disease Detection*, Proc. 3rd Int. Symp. on Artificial Intelligence for Digital Transformation, ACM, 2022. (ACM DL) ACM Digital Library+1

[3] H. Heidari, G. Hellstern and M. Murugappan, "Heart Disease Detection using Quantum Computing and Partitioned Random Forest Methods," arXiv:2208.08882, Aug. 2022. arXiv

[4] A. Surya, "Machine Learning and Ensemble Approach Onto Predicting Heart Disease," arXiv:2111.08667, Nov. 2021. arXiv

[5] *Random forest swarm optimization-based for heart diseases diagnosis*, (conference/journal item), 2021. (see conference/journal listing). ScienceDirectAstrophysics Data System

[6] *Kernel random forest with black hole optimization for heart diseases*, PeerJ Computer Science (article page). 2023/2022 related work. PeerJ

[7] *Supervised Machine Learning-Based Cardiovascular Disease Prediction*, Wiley / Mathematical Biosciences and Engineering, 2021. Wiley Online Library

[8] A. Juliette Albert, "Diagnosis of heart disease using oversampling methods and decision tree classifier in cardiology," *[Journal / repository]*, 2022. (PMC listing). PMC

[9] *Predicting Heart Disease using Random Forest*, ADS / JPhCS abstract (2021). Astrophysics Data System

[10] (Collection of optimized-RF / hybrid RF implementations for heart disease detection — representative examples across 2021–2022). See review and examples in SpringerLinkNature

[11] K. Kwakye and E. Dadzie, "Machine Learning-Based Classification Algorithms for the Prediction of Coronary Heart Diseases," *arXiv preprint arXiv:2112.01503*, 2021.

[12] K. Subramanian, S. Rajasekaran, and R. A. Devi, "Risk Factor Prediction by Naive Bayes Classifier, Logistic Regression Models, Various Classification Techniques," *Proc. Natl. Acad. Sci., India, Sect. B Biol. Sci.*, vol. 92, no. 3, pp. 469–477, 2022.

[13] A. Surya, "Machine Learning and Ensemble Approach Onto Predicting Heart Disease," *arXiv preprint arXiv:2111.08667*, 2021.

[14] S. Kumar et al., "Heart Disease Prediction using distinct artificial intelligence techniques: performance analysis and comparison," *J. Healthcare Eng.*, Article ID 8658763, 2021.

[15] M. A. Javeed et al., "A hybrid AI approach for cardiovascular prediction," *Biomed Res. Int.*, vol. 2021, Article ID 1129428, 2021.

[16] T. G. R. S. Narayana and N. Nalini, "Prediction of fetal heart disease detection using naïve bayes classifier and comparing with linear regression classifier," in *Proc. Int. Conf. on Advances in Data-driven Computing and Predictive Analytics (ADDEPC)*, AIP Conf. Proc., vol. 2853, no. 020012, pp. 1–6, 2021 (published 2024).

[17] A. Saxena, P. Jankisharan, V. S. Kushwah, and A. Mishra, "Prediction of Heart Disease using Machine Learning Algorithms," *ResearchGate Preprint*, 2021.

[18] S. Saraswat, S. Gabhane, A. Pawar, and S. Patil, "Heart Disease Prediction Using Classification (Naïve Bayes)," *ResearchGate Publication*, 2023.

[19] M. M. Ahsan and Z. Siddique, "Machine Learning-Based Heart Disease Diagnosis: A Systematic Literature Review," *arXiv preprint arXiv:2112.06459*, 2021.

[20] P. K. Sahoo et al., "Computational Learning Model for Prediction of Heart Disease Using Machine Learning Based on a New Regularizer," *Front. Cardiovasc. Med.*, vol. 8, pp. 1–10, 2021.

[21] Y. Khourdifi and M. Bahaj, "A decision support system for heart disease prediction based upon SVM, KNN, MLP, RF, and NB classifiers optimized by ACO and PSO," *Int. J. Intell. Eng. Syst.*, vol. 12, no. 1, pp. 242–252, 2021.

[22] P. K. Bhunia, A. Debnath, P. Mondal, M. D. E, K. Ganguly, and P. Rakshit, "Heart Disease Prediction using Machine Learning," *Int. J. Eng. Res. Technol.*, vol. 9, no. 11, pp. 1–6, 2021.

[23] M. Mijwil, "Prediction of heart diseases utilising support vector machine and artificial neural network," *ResearchGate*, 2023. [Online]. Available:

[24] A. Thomas and J. Jyothirmayi, "Early and precise identification and diagnosis of heart disease using support vector machine compared with back-propagation neural network algorithm," *AIP Conf. Proc.*, vol. 3193, no. 1, p. 020087, 2024.

[25] P. K. Bhunia, A. Debnath, P. Mondal, M. D. E, K. Ganguly, and P. Rakshit, "Heart Disease Prediction using Artificial Intelligence," *Int. J. Eng. Res. Technol.*, vol. 9, no. 11, pp. 1–6, 2021.

[26] A. Lakshmanarao, A. Srisaila, and T. S. R. Kiran, "Heart Disease Prediction using Feature Selection and Ensemble Learning Techniques," in *Proc. 2021 Third Int. Conf. Intell. Commun. Technol. Virtual Mobile Networks*, 2021, pp. 994–998.

[27] M. Elsedimy, A. M. El-Sayed, and M. A. El-Bakry, "New cardiovascular disease prediction approach using support vector machine and quantum-behaved particle swarm optimization," *Multimedia Tools and Applications*, vol. 82, no. 3, pp. 1–21, 2023. SpringerLink+1

[28] A. Shrivastav, S. K. Shrivastav, and A. K. Sharma, "Prediction of Heart Disease and Survivability using Support Vector Machine and Naive Bayes Algorithm," *bioRxiv*, 2023. ResearchGate+1
[29] N. M. Lutimath, S. S. Patil, and S. S. Patil, "Classification models combined with Boruta feature selection for heart disease prediction," *Materials Today: Proceedings*, vol. 72, pp. 1–5, 2023. ScienceDirect+1
[30] M. Padilla Rodriguez and M. Nafea, "Centralized and Federated Heart Disease Classification Models Using UCI Dataset and their Shapley-value Based Interpretability," *arXiv preprint arXiv:2408.06183*, 2024. arXiv+1