# Hybrid Machine Learning based Heart Disease Detection

Kuldeep Vayadande, Kunal Sonne, Hariom Surwase, Atharva Tuljapurkar,
Shreeparth torawane , Raj Thakur,
Department of Information and Technology,
Vishwakarma Institute of Technology,
Upper Indra Nagar, Bibewewadi, Pune, India.

***Abstract* —** Accurate and non-deceptive forecast of cardiovascular disease (CVD) may be the most important factor in enhancing clinical outcomes and keeping timely interventions on track. In recent years, ensembling have been highly effective in predictive modeling have been developed by combining the various different base learners to create more accurate and generalized models. Among these, stacking - a technique that integrates the prediction of various classifiers using a final classifier (meta learner) has shown significant promise. This study delves into a comparative analysis of various predictive algorithms such as Logistic Regressor, Nearest Neighbor Classifier, Decision Trees, and Randomized Forest techniques are utilized in data science against a sophisticated stacking ensemble framework. The proposed model employs a combination of RF, DT, and KNN as base learners, while a Ridge Classifier acts as the meta-learner to aggregate and fine-tune the predictions. Through a rigorous evaluation on a publicly available heart.csv dataset, the Random Forest model has a notable accuracy of 88.52%, standing out as the best-performing single algorithm. However, the stacking ensemble outperformed all individual models, achieving an improved accuracy of 90.16%. The outstanding execution of stacking ensemble can be referred to its ability to reduce overfitting, mitigate bias, and capture complex data patterns. These results emphasize the potential of using diverse base models and a strategically chosen meta-learner to enhance predictive performance. The study provides valuable insights into designing optimized ensemble models for clinical risk assessment, ultimately contributing to better decision support and patient management in CVD diagnosis.

***Keywords*** *— Stacking Ensemble, Random Forest, KNN, CVD, Ridge Classifier, Classification, ML models.*

## I. INTRODUCTION

Cardiovascular diseases (CVDs) leads to almost 17 to 18M (in millions) deaths per year [1] . Early and proper detection of CVD by taking an ECG is a key strategy in the prevention of the associated risks and mortality rates. Conventional procedures for diagnosis, despite their suitability, have difficulty avoiding invasive interventions with high fees and receiving late results, thus causing delays in appropriate treatment [4]. Consequently, there is an urgent demand for better, faster, and easier-to-use diagnostic methods which can facilitate the detection and active treatment of CVDs. [6].
Recent advancements in machine learning (ML) have transformed various industries, including healthcare. ML algorithms are proficient in analysing large datasets, recognizing complex patterns, and making highly accurate predictions [5]. In the existing era of CVDs detection, ML provides a promising alternative to traditional methods by utilizing data-driven approaches to predict the disease's onset at an early stage [3]. This capability is particularly important as it facilitates early intervention, lifestyle changes, and treatment adjustments, significantly enhancing patient outcomes [9].

This study plans to construct a machine learning-based mechanism and its validation for diagnosing heart diseases at the early stage. The suggested system aims to achieve the desired high predictive accuracy by utilizing a mixture of clinical and demographic data [2]. The method employs supervised learning algorithms that are trained on large datasets to reveal the factors and early signs of CVDs. The unique feature of the system is its capacity to process as well as analyze complex patterns in real time. This fact gives the potential of making the clinical workflows more efficient and minimizing the medical workers' burden.[7].
This study addresses the task of various ML algorithms and then builds a stacked ensemble out of them. The hybrid Stacking Classifier is a composition of RF, KNN, and Decision Tree as base estimators had the highest score among all individual models with an accuracy of 90.16% [8]. This indicates the effectiveness of hybrid models in enhancing diagnostic precision, and therefore, they can be applied in clinical decision-making [5].

## II. LITERATURE REVIEW

**[11] Paper 1:** "Heart Disease Prediction Using Stacking Ensemble Model"

This paper proposes an innovative approach by leveraging a stacking-based ensemble method. The model incorporates RF, XGBoost, and DT classifiers to predict, having an accuracy score of 88.92% on data set. They highlight the effectiveness of stacking in enhancing overall prediction by combining multiple classifiers. The research emphasizes that feature selection plays an dominant role in further refining model score or accuracy, as selecting important features ensures that model captures some key insights and recurring patterns that are observed in the given data.

**[12]Paper 2 :** "Optimized Ensemble Learning Approach with Explainable AI"

In this paper the authors introduce a refined ensemble model combining Random Forest, XGBoost, and AdaBoost, optimized through Bayesian methods for hyperparameter tuning. The trained model achieved a notable score (accuracy) of 90% on data-set. This research also has the integration of SHAP values (Explainable AI) to interpret model predictions. This particular feature is of utmost importance to the clinical decision-making process as it not only provides transparency and insight into predictions that the proposed model makes and also enables the healthcare professionals to comprehend and trust the model's outputs.

**[13] Paper 3:** "ML for Heart disease predictions with Stacking Techniques"

This paper employs a stacking ensemble of RF, XGBoost, and SVM. Through hyperparameter tuning, having score/accuracy of 89%. The paper discusses how stacking enhances model performance by effectively handling diverse patient data and ensuring robustness across various datasets. The study underscores the importance of stacking in improving predictive accuracy and managing the variability inherent in medical data.

**[14] Paper 4:**"Improving Cardiovascular Disease Detection Using Stacking and Feature Selection"

This paper explores the benefits of stacking ensembles combined with feature selection techniques. The study demonstrates that a stacking ensemble of Random Forest and XGBoost can outperform individual classifiers, achieving 90% accuracy. The authors focus on the impact of selecting relevant features as age, thalach, cholesterol levels, and resting and fasting BP on model performance. Their findings suggest that effective feature selection, coupled with stacking, enhances the model's generalization capabilities and predictive accuracy.

**[15]Paper 5:** "Ensemble Model-Based Prediction of heart diseases."

Paper investigates use of ensemble techniques involving RF, XGBoost, and KNN. The proposed stacking approach achieved a better accuracy of 91% on Framingham data-set. This study highlights how different ensemble methods can manage noisy or incomplete data, which is often encountered in CVDs prediction. The research demonstrates the efficacy of combining multiple ensemble methods to handle diverse data challenges and improve prediction accuracy.

[16]Paper 6: "Predicting heart diseases using ML and exploring feature selection techniques"

This study presents a comparison on methods of selection of features used in combination with ML algorithms for predicting. The research employs the UCI dataset and includes methods for determining the most important features of a heart. These chosen features were then used as the basis to train SVM, GB, and DT classifiers. The RFE-SVM combination yielded the highest accuracy of 89%. This research demonstrates that appropriate selection is important for performance, overfitting avoidance, and ability to interpret the heart disease prediction.

## III. METHODOLOGY

This paper focuses on heart disease detection via various machine learning models. The dataset employed in this project consists of data relevant to CVDs detection. The subsequent parts of this article summarize the whole data preparation, model training, evaluation, and model comparison process. *Fig.2* gives a broad idea about the system architecture.

Data processing for the prediction model was a thorough process done in Python using the pandas library, first followed by the feature scaling with sklearn. The main goal was to tune the dataset for a different kind of machine learning algorithms especially distance-based methods such as KNN and algorithms that are responsive to feature magnitude like Logistic Regression. Below is an in-depth explanation of the methodology followed:

**Data Loading and Inspection:** The heart disease dataset, typically in CSV format, was imported using the pandas library. An initial examination was conducted to understand the structure and attributes of the dataset. Key steps included checking for the presence of null values, understanding data types, and analysing basic statistics. This analysis helped in determining which features were relevant and which ones needed preprocessing.

**Standardization of Features:** Feature scaling is a critical step, especially when dealing with algorithms that rely on distance calculations, such as KNN, or models that assume normal distribution, like Logistic Regression. Standardization was applied using class StandardScaler from module sklearn.preprocessing. This technique transformed every feature , so that it had a mean equals to zero and Standard Deviation equal to one. By standardizing, all features were brought to the same scale, eliminating

the risk of higher magnitude variables dominating the learning process. This was particularly important for optimizing the performance of KNN, which is sensitive to the scale of the input data, and Logistic Regression, which benefits from standardized features for faster and more reliable convergence.

**Splitting the Dataset:** Initially, the dataset which was pre-processed was divided in two Training and Testing subsets. The data was then split using the typical 80-20 split ratio, where 70% to 80% data was reserved for training of model, and some part of that data which remained was later on utilized for testing and assessment. The separation was all the same, training and taking tests the model on different data, so it freed overfitting and made it capable to generalize on unseen(Testing) data. Then data was split in such a manner where 80% was consumed by training and 20% for purposes of testing. After splitting the dataset, the subsequent machine learning models were used: KNN:This algorithm was applied with 7 neighbors and the test data was 83% accurate. This model evaluated the heart disease risk depending on the closeness of the data points. Decision Tree: Using random states, the Decision Tree model was trained on a dataset with an accuracy of 80%. It is the method that has a strong point in the handling relationships that are non-linear. RF Classifier: After training with several random states, the Random Forest model obtained best accuracy score among individual models, with a score of 88.16%. The ensemble model (RF is itself an ensemble model of decision trees) is the reason for its better accuracy and lower overfitting. *Fig.3* gives an overview of how RF algorithm is an ensemble model. *Fig .1* explains the Proposed ensemble model(stacking classifier).


**Fig. 1:** Proposed Ensemble Model

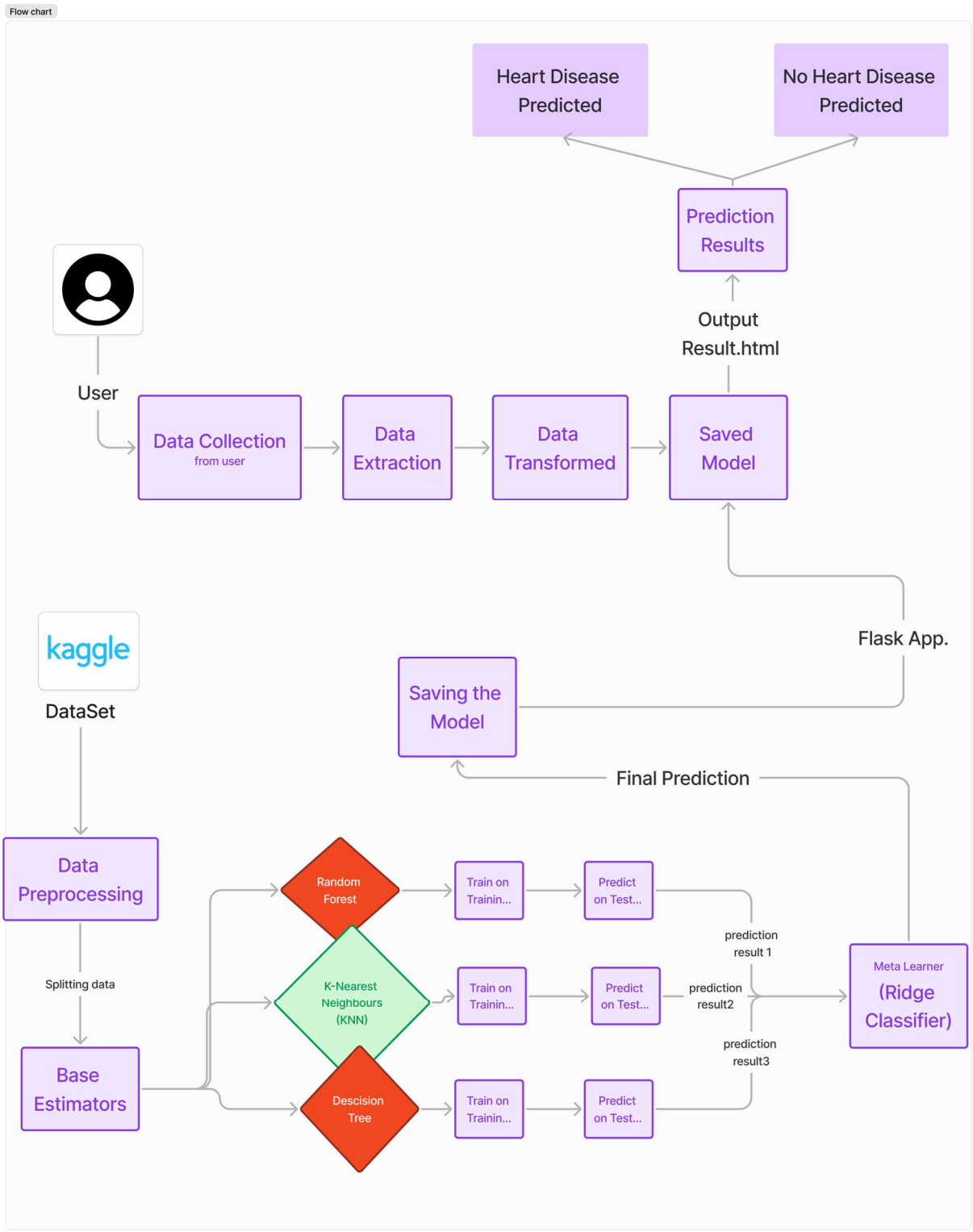


**Fig. 2:** Proposed Flow Diagram

**Table 1:** Formulae Table

| **Statistical Metrics to Evaluate Prediction Model Performances** | |
|---|---|
| ACCURACY | $\frac{T.P + T.N}{T.P + TN + F.P + FN}$ <br> TP if true positive, TN is true negative, FP is false positive, and FN is false negative |
| RECALL | $\frac{T.P}{T.P + F.N}$ |
| PRECISION | $\frac{T.P}{T.P + F.P}$ |
| F1- SCORE | $\frac{2 * RECALL * PRECISION}{RECALL + PRECISION}$ |
| Area Under ROC curve / AUC | $F.P.R = \frac{F.P}{F.P+T.N}$ |
| | $T.P.R = \frac{T.P}{T.P+F.N}$ |
| | $AUC = \int TPR(FPR) dFPR$ |

**Algorithm 1: KNN**

In heart disease prediction, KNN can provide a straightforward way to classify individuals based on the similarity of their health features to those of known cases**.** Calculation of distance: First we have to calculate distance between the instances which are new and instances in training set.

Neighbour Selection: Select the k closest instances (neighbours).

Majority Voting (for classification): Assign most common class in k neighbors.

Weighted distance-

$$\sqrt{\sum_{i=1}^{n} wi\,(xi - yi)^2}$$

**Algorithm 2:** Decision Tree

For heart disease detection, Decision Trees can clearly show how features like age and cholesterol levels are used to make decisions about the likelihood of heart disease.

1. **Splitting:** At every node, data is first split which is based on feature resulting the most information gained or can be the on with lowest impurity.
2. **Stopping Criteria:** Splitting stops when a leaf node is reached or when additional splits no longer improve the model.

**Information Gain:**

$$\boldsymbol{IG = Entropy(parent) - \sum_{i=1}^{k} \frac{|Di|}{|D|}\ Entropy(Di)}$$

*|D| is the size of the parent dataset, and |Di| represents subsets after the split.*

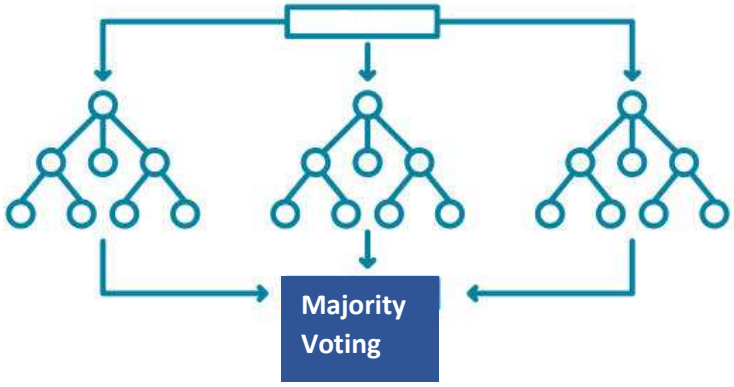


**Fig. 3:** Random Forest uses collection of decision trees

**Algorithm 3:** Random Forest

In prediction, Random Forest is able to handle complex interactions between features (e.g., age, cholesterol levels, etc.) and improve classification accuracy.

1. **Bootstrapping:** Random subsets of the training data are used to build each decision tree.
2. **Feature Randomness:** On every successful split in trees, a randomized subset consisting of certain features is taken.

$$Gini = 1 - \sum_{i=0}^{k} (pi)^2$$

$$OOB\ error = \frac{1}{n_{OOB}} \sum_{i \in OOB}^{k} Indicator(\hat{y}_i \neq y_i)$$

*where $n_{OOB}$ is the number of out−of−bag samples, $\hat{y}_i$ is the predicted value, and $y_i$ is the true value. $pi$ is probability of an element being classified into class $i$.*

| Model | Precision | Recall | F1-score |
|---|---|---|---|
| Stacking Model | 92 | 94 | 91 |
| RF | 0.89 | 0.91 | 0.90 |
| K-NN | 0.84 | 0.78 | 0.81 |
| DT | 0.84 | 0.79 | 0.82 |

**Table 2:** Comparison of Models w.r.t to Precision,Recall,F1-score

| Model | ACCURACY |
|---|---|
| Stacking Model | 90.16% |
| Random Forest | 88.52% |
| KNN | 83.16% |
| Decision Tree | 80.12% |

**Table 3:** Comparison of models w.r.t accuracy

| Paper | ACCURACY |
|---|---|
| Proposed Stacking Model | 90.16% |
| [11] | 88.92% |
| [12] | 90.0% |
| [14] | 89% |
| [15] | 90% |
| [16] | 91.12% |

Table 4: Proposed Model comparison with existing models

## IV. Key-Metrics

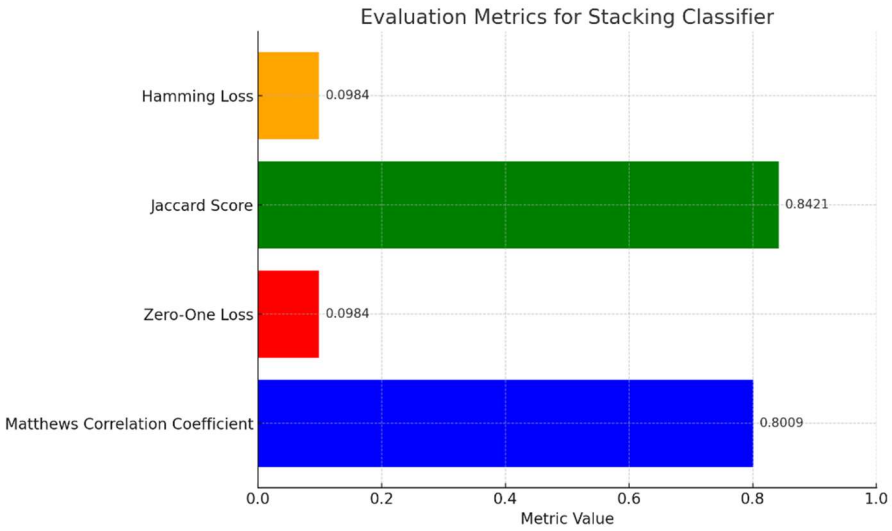


Figure. 3: Evaluation metrics

**MCC (0.8009)** indicates strong predictive capability, especially useful for imbalanced data . **Zero-One Loss (0.0984)** and **Hamming Loss (0.0984)** show that our model is making less than 10% errors. **Jaccard Score (0.8421)** confirms that the model captures positive class labels well, with around 84% similarity between actual and predicted positive instances. *Fig.3 and Fig.4* shows all metrics plotted using pandas library in python.

| Metric | PRECISION | RECALL | F1-SCORE | SUPPORT |
|---|---|---|---|---|
| Class indication 0 | 0.92 | 0.85 | 0.88 | 27 |
| Class indication 1 | 0.89 | 0.94 | 0.91 | 34 |
| ACCURACY | - | - | 0.90 | 61 |
| MACRO AVERAGE | 0.90 | 0.90 | 0.90 | 61 |
| WEIGHTED AVERAGE | 0.90 | 0.90 | 0.90 | 61 |

Table. 5: Stacking Model Metrics

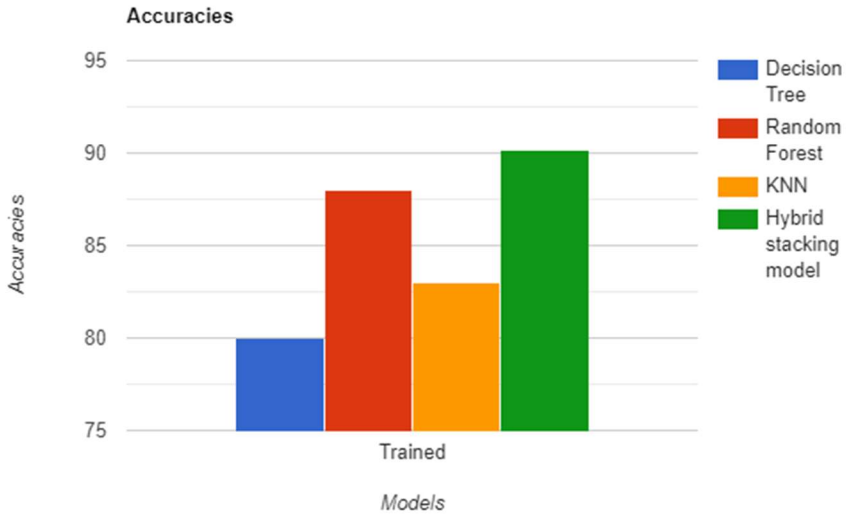


Figure. 4: Accuracy plot

## V. RESULTS & CONCLUSION

In this research, a model was developed for predicting heart diseases. The proposed approach hybrid Stacking Classifier with Ridge Classifier as its Meta Learner. The benefit of proposed approach is demonstrated by the achieved score /accuracy of 90.16% and AUC value of 0.97 (shown in *Table.2* and *Table.3*). This demonstrates that hybrid models outperform individual classifiers by utilizing the complementary strengths of different models. The use of ensemble techniques such as Stacking Classifier shows clear improvement, providing better diagnostic performance, which is crucial in clinical settings where accuracy is critical for decision-making. The achieved accuracy of 93% indicates that the proposed approach is effective for predicting CVD. However, further research can be conducted to explore the potential of using stacking hybrid models and optimal techniques to further enhance the prediction score or accuracy of the model. Overall, the proposed approach provides a promising solution for predicting the heart diseases. The proposed approach provides better performance than the traditional ML based model. The use hybrid stacking classifier enhances the model performance which are shown in Table 3.

## VI. FUTURE SCOPE

In future, this research can be expanded to make heart disease detection more accessible and user-friendly. One key area of development is incorporating data from wearable devices and IoT systems. This would allow for continuous, real-time health monitoring, where devices like smartwatches could track heart activity and provide warnings about risks. Additionally, more advanced machine learning models, such as deep learning or other hybrid algorithms, could be explored to improve prediction accuracy. By integrating these models with larger and more diverse datasets such as genetic data, imaging, or patient history we can move toward more personalized predictions, made to individual patients' unique medical profiles. Another exciting direction is the potential for integrating these tools into clinical environments, providing healthcare professionals with an assistant to support diagnosis and treatment decisions.

## VII. REFERENCES


[1] Mythili, T., et al. "A heart disease prediction model using SVM Decision Trees Logistic Regression (SDL)." International Journal of Computer Applications (2013).

[2] Mohan, Senthilkumar, Chandrasekar Thirumalai, and Gautam Srivastava. "Effective heart disease prediction using hybrid machine learning techniques." IEEE Access (2019).

[3] Singh, A., et.al (2020, February). Heart Disease Prediction Using Machine Learning Algorithms. 2020 International Conference on Electrical and Electronics Engineering (ICE3) IEEE.(450-455).

[4] Jonnavithula, et.al (2020). Role of machine learning algorithms over heart diseases prediction. In AIP Conference Proceedings AIP Publishing LLC. (No. 1,p.040013).

[5] Seral Şahan ,Polat, Kemal, and Salih Güneş. "Automatic detection of heart disease using an artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism and k-nn (nearest neighbour) based weighting preprocessing." Expert Systems with Applications 32.2 (2007).

[6] S. Swamy and N. K. S. Banu ‘‘Prediction of heart disease at early stage using data mining and big data analytics: A survey,’’ in Proc. Int. Conf. Elect., Electron., Commun., Comput. Optim. Techn. (ICEECCOT), Dec. 2016.

[7] I. Turkoglu, A. Sengur and R. Das, ‘‘Effective diagnosis of heart disease through neural networks ensembles,’’ Expert Syst. Appl., vol. 36, no. 4, pp. 7673–7679, May.

[8] R. R. Rajalaxmi and A. S. Abdullah , ‘‘A data mining model for predicting the coronary heart disease using random forest classifier,’’ Commun. Controls, Apr. 2012, pp. 20–27.

[9] K. J. Shanthi, D. K. Ravish, , N. R. Shenoy, and S. Nisargh, ‘‘Heart function monitoring, prediction and prevention of heart attacks: Using artificial neural networks,’’ Comput. Inform. (IC3I), Nov. 2014, pp. 3-9.

[10] P. Sumitra and C. Sowmiya ‘‘Analytical study of heart disease diagnosis using classification techniques,’’ in Proc. IEEE Optim. Signal Process. (INCOS), Mar. 2017, pp. 3–9.

[11] N. N. Itoo and V. K. Garg, "Heart Disease Prediction using a Stacked Ensemble of Supervised Machine Learning Classifiers," *2022 International Mobile and Embedded Technology Conference (MECON)*, Noida, India, 2022, pp. 599-604, doi: 10.1109/MECON53876.2022.9751883.

[12] Mienye, Domor & Jere, Nobert. (2024). Optimized Ensemble Learning Approach with Explainable AI for Improved Heart Disease Prediction. Information. 15. 394. 10.3390/info15070394.

[13] Subasish Mohapatra, Sushree Maneesha, Prashanta Kumar Patra, Subhadarshini Mohanty, Heart Diseases Prediction based on Stacking Classifiers Model (Procedia Computer

Science,Vol. 218) 2023 ISSN 1877-0509,https://doi.org/10.1016/j.procs.2023.01.140, pp. 1621-1630.

[14] Chowdary, K Rohit et al. “Heart Disease Prediction using Feature Selection and Ensemble Learning Techniques.” *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)* (2021): 994-998.

[15] Y. Shaikh, V. K. Parvati and S. R. Biradar, "Heart Disease Prediction using Ensemble Learning," *2023 International Conference on Applied Intelligence and Sustainable Computing (ICAISC)*, Dharwad, India, 2023, pp. 1-5.

[16] Mahmudul Hasan, Md Abdus Sahid, Md Palash Uddin, Md Abu Marjan, Seifedine Kadry, Jungeun Kim, Performance discrepancy mitigation in heart disease prediction for multisensory inter-datasets, PeerJ Computer Science, 10.7717/peerj-cs.1917, **10**, (e1917), (2024).