IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, February 2024

# Enhancing Heart Disease Prediction using Advanced Feature Engineering and Ensemble Learning Techniques

**Chandana C, Bhavya Sree S, Prof. Mahendra MK**
Department of Information Science and Engineering
Global Academy of Technology, Bengaluru, India

**Abstract:** *This study introduces a holistic model for predicting heart disease ,integrating advanced algorithms with a focus on feature engineering. The dataset encompasses a diverse range of patient parameters, including demographics, lifestyle factors, and medical history.*

*Feature engineering involves a meticulous process of selecting, transforming, and augmenting relevant features to enhance the model's ability to discern patterns and relationships within the data. This stage is essential for increasing the predicted accuracy of the model and gleaning insightful information from intricate datasets.*

*The logistic regression algorithm is employed to establish a baseline predictive model, providing insights into the individual contribution of each feature. Subsequently, a neural network is implemented to capture intricate non-linear dependencies and interactions within the data, further refining the predictive capabilities.*

*Results indicate that the incorporation of feature engineering significantly improves the model's performance compared to traditional approaches. Early experiments demonstrate promising accuracy rates in heart disease prediction, showcasing the potential for early detection and proactive healthcare interventions.*

*This method not only enhances predictive modeling for heart disease but also emphasizes the significance of feature engineering in maximizing the full capabilities of machine learning algorithms for medical applications.*

**Keywords:** Heart disease prediction, feature engineering, logistic regression, neural network, dataset, predictive accuracy, machine learning, early detection, healthcare interventions, medical applications

## I. INTRODUCTION

This research endeavors to enhance the predictive capabilities of heart disease models by employing a sophisticated blend of advanced feature engineering and ensemble learning techniques. Recognizing the multifaceted nature of cardiovascular health, we harness a rich and diverse dataset comprising demographic nuances, lifestyle intricacies, and detailed medical histories.

Feature engineering takes center stage in our methodology, involving a meticulous process of selecting, transforming, and augmenting features. This strategic refinement aims to unveil subtle patterns within the data, empowering the model to discern intricate relationships crucial for precise prediction of heart disease.

In tandem with feature engineering, we embrace the power of ensemble learning. This involves the amalgamation of diverse algorithms, each contributing its unique strengths to create a more robust and resilient predictive framework. Ensemble learning not only elevates predictive accuracy but also enhances the model's adaptability to the complexities and uncertainties inherent in healthcare datasets.

Our research is not merely an academic pursuit but a practical endeavor with far-reaching implications. By combining advanced feature engineering with ensemble learning, we aspire to furnish a predictive model that not just excels in accuracy but also serves as a proactive tool for healthcare practitioners, facilitating early intervention and preventive strategies in the realm of heart disease.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, February 2024

## II. LITERATURE SURVEY

In [1],This paper, published in IEEE Access in May 2023, presents a novel Principal Component Heart Failure (PCHF) feature engineering approach for heart disease prediction. The authors employed nine machine learning-based algorithms and proposed the PCHF mechanism to select the most prominent features to enhance performance. They optimized the PCHF mechanism by creating a new feature set to optimalthe highest accuracy scores. The newly created dataset is based on the eight best-fit features. The proposed decision tree method outperformed the applied machine learning models and other state-of-the-art studies, achieving a high accuracy score of 100%. This study contributes significantly to the medical community by improving early detection of heart failure [1].

In [2], The suggested model demonstrated superior performance in accuracy, sensitivity, and specificity compared to alternative techniques. Moving forward, we aim to enhance this approach by incorporating image data from patients with heart disease. These images will be collected through laboratory examinations and imaging techniques. Additionally, we plan to employ Convolutional Neural Networks (CNNs) to diagnose heart disease with optimal accuracy.

In [3] In a 2021 publication featured in Computational Intelligence and Neuroscience, the authors explore the application of diverse machine learning algorithms and deep learning techniques on the UCI Machine Learning Heart Disease dataset, which comprises 14 primary attributes for analysis. Through this comparative study, promising results are attained and evaluated using metrics like accuracy and confusion matrix. To address irrelevant features within the dataset, the authors employ Isolation Forest, while also normalizing the data to enhance performance. Employing a deep learning methodology, the authors achieve an accuracy rate of 94.2%.

In [4], Published in the 2022 Advances in Science and Engineering Technology International Conferences, this paper presents several machine learning approaches for predicting heart diseases, using data of major health factors from patients. The paper demonstrated four classification methods: Multilayer Perceptron (MLP), Support Vector Machine (SVM), Random Forest (RF), and Naïve Bayes (NB), to build the prediction models. Data preprocessing and feature selection steps were done before building the models. The models were evaluated based on the accuracy, precision, recall, and F1-score. The SVM model performed best with 91.67% accuracy [4].

In [5],Published in the 2022 Advances in Science and Engineering Technology International Conferences, this paper presents several machine learning approaches for predicting heart diseases, using data of major health factors from patients. The paper demonstrated four classification methods: Multilayer Perceptron (MLP), Support Vector Machine (SVM), Random Forest (RF), and Naïve Bayes (NB), to build the prediction models. Data preprocessing and feature selection steps were done before building the models. The models underwent assessment utilizing metrics including accuracy, precision, recall, and F1-score. The SVM model exhibited superior performance, achieving a 91.67% accuracy rate.

In [6], The presented approach for predicting heart disease employs a decision tree-based random forest (DTRF) classifier with loss optimization. Initially, the dataset undergoes preprocessing, focusing on patient records with known labels indicating the presence or absence of heart disease. Subsequently, a DTRF classifier is trained using a stochastic gradient boosting (SGB) loss optimization technique, and its performance is assessed on a separate test dataset. Results indicate that the proposed HDP-DTRF approach achieves higher metrics, with 86% precision, 86% recall, 85% F1-score, and 96% accuracy on publicly available real-world datasets, surpassing traditional methods [6].

In [7], This paper presents a hybrid machine learning system that combines multiple machine learning algorithms to create a more robust and accurate prediction model for heart disease. The authors argue that while individual machine learning algorithms have their strengths, they also have limitations that can affect the accuracy of their predictions. By combining these algorithms, the hybrid system leverages the strengths of each algorithm while mitigating their weaknesses. The hybrid system was trained and tested on a dataset of patient records, which included various health indicators such as age, sex, blood pressure, cholesterol levels, and more. The system achieved a high accuracy rate, outperforming the individual algorithms. This suggests that a hybrid approach can enhance the prediction of heart disease, potentially leading to earlier interventions and better patient outcomes [7].

In [8], This review provides a comprehensive overview of the applications of deep learning and machine learning in predicting cardiovascular diseases. The authors discuss various machine learning and deep learning models, including decision trees, support vector machines, neural networks, and convolutional neural networks. They highlight the

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, February 2024

advantages of these models, such as their ability to handle large datasets and complex relationships between variables, as well as their limitations, such as the risk of overfitting and the need for large amounts of data. The review also discusses the performance of these models in predicting heart diseases, with some models achieving high accuracy rates. However, the authors note that the effectiveness of these models can vary depending on the dataset and the specific task. The review concludes by highlighting the potential of machine learning and deep learning models in improving patient care and outcomes, but also emphasizes the need for further research to optimize these models and address their limitations [8].

In [9], This paper presents a survey of various models based on bio-inspired algorithms and techniques and analyzes their performance in predicting heart diseases. The authors discuss several machine learning techniques that have been used by researchers to assist the healthcare industry and professionals in diagnosing heart-related diseases1. The paper provides a comprehensive overview of the current state of heart disease prediction models and offers insights into the effectiveness of bio-inspired algorithms in this domain [9].

In[10], Published in the IEEE Xplore, this paper presents a medical system learning-based medical system for predicting the forecastinga heart disease occurrence in patients. The study utilizes the UCI dataset to analyze multiple indicators using eight different algorithms to identify the most accurate and comprehensive attributes for predicting heart disease. The authors conduct a in-depth analysis of each algorithm's performance and discuss the implications of their findings for the advancement of heart disease prediction [10].

In [11],This paper developed and evaluated the Enhanced Deep learning assisted Convolutional Neural Network Learning Prediction Models and Classification, depends on diagnostic performance in diagnostic odds ratio, 95 % confidence interval using the sensitivity and specificity of the heart disease. The enhanced deep learning prediction models and classification has been constructed with a deep multi-layer perception equipped to create a secure and improved classification model with non-linear functions and linear, regularization, and falling and binary sigmoid classifications utilizing dedicated learning technologies.

In [13], The suggested model demonstrated superior performance in accuracy, sensitivity, and specificity compared to alternative techniques. Moving forward, we aim to enhance this approach by incorporating image data from patients with heart disease. These images will be collected through laboratory examinations and imaging techniques. Additionally, we plan to employ Convolutional Neural Networks (CNNs) to diagnose heart disease with optimal accuracy.

In [14], The aim of this project is to forecast heart disease occurrence by analyzing medical histories, physiological measurements, and lifestyle factors. Through model training and evaluation, the study seeks to accurately differentiate Individuals afflicted by and free from heart conditions. Challenges include data quality and model complexity, demanding iterative refinement. This research aligns with the growing need for early detection, intervention, and personalized treatment strategies, ultimately improving cardiovascular healthcare outcomes

, In [15], The study in this paper highlights the transformative potential of big data analytics (BDA) in organizations, emphasizing its crucial role in supporting the circular economy (CE). By empirically examining 109 Czech manufacturing firms, the research establishes a positive association between BDA capability, business intelligence and analytics (BI&A), and decision-making quality, with data-driven insights enhancing this relationship. The findings underscore the importance of BDA capability in driving decision-making quality within the CE paradigm, providing valuable insights for managers in developing data-driven strategies

| SL NO | PAPER TITLE | TECHNIQUES | ADDRESSED ISSUES |
|---|---|---|---|
| 1 | Heart disease data prediction based privacy preservation using enhanced ElGamal and ResNet classifier | The technique used in the context is "logistic regression." | ML models may inheritbiases present in the training data, leading to inaccurate predictions, especially if the data can be utlized is not diverse or representative. |
| 2 | Heart Disease Prediction using"ComputationalIntelligence | Ensembleing learning models like bagging | Leveraging pre-trained models on related tasks to improve the performance of heart |

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, February 2024

| | | | |
|---|---|---|---|
| | and Communication Technology" | and boosting are used | disease prediction models,and has limited data |
| 3 | Early-stage detection of heart failure using machine learning techniques. | The technique used is deep learning models will be trained | limited Labeled Data: Obtaining labeled data for early-stage heart failure is challenging due to the asymptomatic nature of the condition. Limited data can hinder the training of robust and accurate models. |
| 4 | HDPM: an effective heart disease prediction model for a clinical decision support system. | The technique used is PASCEL model to leverage large clinical data," | XGBoost can predict the presence of heart disease but lacks the ability to differentiate between types of heart disease. |
| 5 | Heart-diseases prediction using deep learning neural network model. I | It uses convolution neural networks(CNN) and electrocar digram(ECG) signals | Imbalance data occurred and impacted on performance of prediction |
| 6 | heart disease Using Stacking model With Balancing Techniques and Dimensionality Reduction | It uses long short term memory(LSTM) for sequential data analysis | Top of Form<br>Dimenisonal technique like PCA or LDA can affect the performance model |
| 7 | A RobustHeart Disease Prediction System employoing the blend of Deep Neural Networks | It leverages pre trained deep learning models(on natural images) | Utilizing XGBoost for predicting heart disease presents challenges. |
| 8 | A refined linear model combined with a recursion-enhanced random forest (RERF-ILM) used for detecting heart disease on an Internet of Medical Things platform." | Attention maps,saliency to interpret deep learning models | RERF-ILM could overfit and which could limit the ability to generalize new unseen data |
| 9 | An improved deep learning-supported CNN used for predicting heart disease on an Internet of Medical Things platform." | optimization of screening testing strategies | Optmization testing strategies used are complex |
| 10 | Heart disease prediction using RERF-ILM and machine learning | Multi model data fusion is used to integrate patients source data | .Complex models like RERF-ILM can be difficult to interpret. |
| 11 | Ensemble methods of heart disease prediction | Gradient boosting methods (GBM) and ada boost are used to train weak learners | Machine learning models need substantial data for effective training. When data is scarce or biased, it can negatively impact the model's performance. |
| 12 | Prediction of heart disease using a combination of ML and deep learning | Grid search and random search is used to optimize model performance | Overfitting occurs and cannot find unseen data |
| 13 | Early prediction of heart disease with data analysis using supervised learning with stocatic | Stacking(combining predictions from multiple models) is | Failure to addressthe computational complexity of multiple algorithms is a limitations of the proposed model. |

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, February 2024

| | | | |
|---|---|---|---|
| | gradient boosting | used | |
| 14 | A hybrid machine learning system for heart disease prediction | Fine-tune pre trained models transfer learning leverage knowledge from related tasks | Complex models couldn't be interpreted |
| 15 | Deep learning and machine learning in cardio vascular diseases:review | SHAP values,LIME are used to understand and communicate predictions | Data requirements are more and couldn't be applied to huge data |

## III. CONCLUSION

Predicting heart disease using machine learning entails a thorough procedure. It beginsas gathering vast datasets containing diverse patient data, spanning demographic information to clinical markers. Essential preprocessing steps, such as cleaning, normalization, and managing missing values, are crucial to uphold the dataset's integrity.

Feature engineering is vital in pinpointing essential variables with substantial influence on heart disease prediction. This may entail leveraging domain-specific expertise or employing methods like recursive feature elimination. In the model selection stage, one must opt for a suitable algorithm—common options being decision trees, SVM, and neural networks—tailored to the intricacies of the data.

Throughout the training process, the model assimilates patterns and connections present within the dataset. Fine-tuning parameters and validating the model ensure optimal performance. Cross-validation techniques assess the model's generalizability. Feature importance analysis sheds light on the variables that carry the most weight in predicting heart disease.

After training and validation, the model undergoes thorough testing on unseen data to evaluate its real-world predictive abilities. Performance metrics like accuracy, precision, recall, and F1 score provide a comprehensive assessment.

Implementing such machine learning systems in clinical settings empowers healthcare professionals with early risk assessment tools. Proactive interventions, tailored to an individual's predicted risk, can range from lifestyle modifications to targeted medical treatments. Continuous monitoring and updates for model ensure adaptability to evolving health scenarios, making it a valuable asset in the ongoing battle against heart diseases.

## REFERENCES

[1] Goyal S (2022) FOFS: firefly optimization for feature selection to predict fault-prone software modules. In: Nanda P, Verma VK, Srivastava S, Gupta RK, Mazumdar AP (eds) Data engineering for smart systems. Lecture Notes in Networks and Systems, vol 238. Springer, Singapore. https://doi.org/10.1007/978-981-16-2641-8_46

[2] Goyal S (2023) Software measurements with machine learning techniques-a review. Recent Adv Comput Sci Commun 16:1–17. https://dx.doi.org/10.2174/2666255815666220407101922

[3] Goyal S (2022) FOFS: firefly optimization for feature selection to predict fault-prone software modules. In: Nanda P, Verma VK, Srivastava S, Gupta RK, Mazumdar AP (eds) Data engineering for smart systems. Lecture Notes in Networks and Systems, vol 238. Springer, Singapore. https://doi.org/10.1007/978-981-16-2641-8_46

[4] Benhar Charles V, Surendran D, SureshKumar A (2022) Heart disease data based privacy preservation using enhanced ElGamal and ResNet classifier. Biomed Signal Process Control 71(Part B):103185, ISSN 1746-8094. https://doi.org/10.1016/j.bspc.2021.103185

[5] Sayad AT, Halkarnikar PP. Diagnosis of heart disease using neural network approach. *Int J Adv Sci Eng Technol.* 2014;**2**:88–92. [Google Scholar]

[6] (2022) Genetic evolution-based feature selection for software defect prediction using SVMs. J Circuits Syst Comput 31(11):2250161. https://doi.org/10.1142/S0218126622501614

[7] Handling class-imbalance with KNN (neighbourhood) under-sampling for software defectprediction. Artif Intell Rev. https://doi.org/10.1007/s10462-021-10044-w

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, February 2024

[8] Predicting the defects using stacked ensemble learner with filtered dataset. Autom SoftwEng28:14. https://doi.org/10.1007/s10515-021-00285-y.
[9] Static code metrics-based deep learning architecture for software fault prediction. Soft Computepp 13. https://doi.org/10.1007/s00500-022-07365-5
[10] Effective software defect prediction using support vector machines (SVMs). Int J Syst AssurEngManag. https://doi.org/10.1007/s13198-021-01326-1
[11] Aakash Chauhan , Aditya Jain , Purushottam Sharma , Vikas Deep, "Heart Disease Prediction using Evolutionary Rule Learning", "International Conference on "ComputationalIntelligence and Communication Technology" (CICT 2020).
[12] Gour, S., Panwar, P., Dwivedi, D. & Mali, C. A machine learning approach for heart attack prediction. Intell. Sustain. Syst. 2555(1), 741–747 (2022).
[13] Juhola, M. et al. Data analytics for cardiac diseases. Comput. Biol. Med. 142(1), 1–9 (2022).
[14] Alom, Z. et al. Early-stage detection of heart failure using machine learning techniques. Proc. Int. Conf. Big Data IoT Mach. Learn. 95, 75–88 (2021).
[15]. Sharma, S. & Parmar, M. Heart-diseases prediction using deep learning neural network model. Int. J. Innov. Technol. Explor. Eng. 9(3), 2244–2248 (2020).