ISSN 2658-5014 (Print) SIIT ISSN 2686-7044 (Online)
2025. Vol. 7, no. 3 (22), pp. 48–65 http://siit.ugatu.su
SYSTEMS ENGINEERING AND INFORMATION TECHNOLOGIES
Scientific paper

UDC 621.9.047 DOI 10.54708/2658-5014-SIIT-2025-no3-p48 EDN IZXXUE

A Novel Paradigm in Cardiovascular Disease Risk Prediction Through Hybrid Machine Learning

PARVEZ RAHI • SANDEEP SINGH KANG

Abstract. Heart disease is known to kill the most in the entire universe, causing deaths of above 17.9 million annually. Early and accurate risk prediction is deemed essential for better clinical outcomes as well as reduced health care burden. This paper proposes an innovative hybrid framework of machine learning that predicts heart diseases with a good degree of accuracy using vital medical as well as lifestyle factors. Such clinically relevant parameters as BMI, diabetic history, hypertensive condition, history of stroke, chronic kidney disease, physical inactivity, and mental disorders by themselves are known well as risk factors in cardiovascular pathology. The hybrid model uses XGBoost, which combines the advantages of both algorithms, SVM and DNN. These advanced engineering methods capture complicated, non-linear correlations between risk variables like diabetes and obesity through polynomial transformations and interaction terms. The SMOTE algorithm helped in classifying the work to alleviate class imbalance and increase prediction accuracy by using a properly balanced dataset to train the model. The suggested method performed better than traditional prediction models, with 94% accuracy. No Risk, Low Risk, Moderate Risk, High Risk, and Severe Heart Disease are the five categories that are used to accurately categorize the risk of heart disease. Four key predictors of heart disease-the algorithm used identified BMI, hypertension, diabetes, and physical health-collated well with current medical understanding. This algorithm represents a powerful tool for clinicians who can use it to stratify their patients on a personal basis and especially identify at an early date those who are at high risk. The model will help healthcare practitioners offer specific treatments by being integrated into clinical practices, thereby eventually resulting in improved outcomes for patients and reduced prevalence of cardiovascular events over time.

Keywords: Cardiovascular disease, BMI, Diabetes, Hypertension, XGBoost, Deep Neural Networks, Risk Stratification, Heart disease prediction, Clinical decision-making.

INTRODUCTION

As the leading cause of morbidity and mortality worldwide, cardiovascular diseases (CVDs) continue to pose a significant threat to public health. The World Health Organization (WHO) estimates that cardiovascular diseases (CVDs) account for nearly 33% of all fatalities annually, or 17.9 million deaths [DiC24]. Of these, ischemic heart disease continues to be the most common kind, and the pathophysiology of these disorders is greatly influenced by risk factors such obesity, diabetes, hypertension, and hyperlipidemia. Among these, ischemic heart disease remains the most prevalent form, with the pathophysiology of these disorders being greatly influenced by risk factors such obesity, diabetes, hypertension, and hyperlipidemia. People in low- and middle-income nations are disproportionately affected by the burden of heart disease, where access to preventative treatment and early intervention remains limited [Pow21].

The pathophysiological mechanisms underlying heart disease are complex and multifactorial, involving interactions among genetic predisposition, lifestyle factors, and comorbid conditions [But22]. Asymptomatic individuals may harbor significant cardiovascular risk factors, often going undetected until acute clinical events, such as myocardial infarction or heart failure, occur [Thu22]. Two examples of conventional risk assessment methods that have long been essential for estimating the likelihood of cardiovascular events are the Framingham Risk Score and the Reynolds Risk Score [ESC21]. However, these tools rely on a limited selection of demographic and clinical parameters, which may not capture the multifaceted nature of cardiovascular risk in diverse populations [Vis24].

Recommended for publication by the Program Committee of the X International Scientific Conference ITIDS'2024 "Information Technologies for Intelligent Decision Support", Ufa–Baku–Chandigarh, November 12–14, 2024.

A potential approach to improving clinical decision-making and prediction accuracy in cardiovascular risk assessment is the recent integration of artificial intelligence and machine learning [Kum23]. Large and complicated datasets may be analyzed using machine learning algorithms, which can spot complex patterns and correlations that conventional statistical techniques would miss [Shu23]. By considering the non-linear interactions between different risk variables, hybrid machine learning models—which incorporate the advantages of numerous algorithms—offer the potential to increase the accuracy of cardiovascular risk predictions [Naz24].

This study proposes a sophisticated hybrid machine learning framework that amalgamates XGBoost, Support Vector Machines (SVM) and Deep Neural Networks (DNN) to facilitate the accurate prediction of heart disease risk [Gha24]. Our model uses a large dataset that includes important clinical characteristics such as BMI, diabetes, hypertension, stroke history, renal function, and important lifestyle variables including physical activity and mental health [Pan20]. To efficiently describe the intricate interactions between these variables, sophisticated feature engineering approaches are used, such as the creation of interaction terms and polynomial transformations [Jia22]. The main objective of this study is to achieve a high level of predictive accuracy, enabling healthcare professionals to identify at-risk individuals earlier in their clinical journey [Moh22]. By implementing this hybrid model within clinical practice, we anticipate enhancing patient outcomes through personalized risk stratification and targeted interventions [Pas20]. The findings from this study are expected to contribute valuable insights into the interplay between various risk factors for heart disease, guiding clinicians in developing tailored prevention strategies [Set23].

The convergence of advanced machine learning methodologies with cardiovascular risk assessment represents a paradigm shift in clinical cardiology [Cha23]. This research endeavors to establish a robust decision-support tool that is not only statistically sound but also clinically relevant, thereby addressing the urgent need for improved cardiovascular disease management and prevention [Paw24]. By enhancing our understanding of heart disease risk factors and their interrelationships, our ultimate goal is to lessen the burden of cardiovascular illness and death by paving the path for more efficient therapies [Sha20].

BACKGROUND AND MOTIVATION

Cardiovascular diseases (CVDs), particularly Over 17.9 million fatalities worldwide are attributed to heart disease each year, making it one of the major causes of death [DiC24]. A rise in risk factors such diabetes, obesity, high blood pressure, and sedentary lifestyles highlights an urgent need for effective risk prediction tools [Bud20]. Accurate risk assessment is vital for facilitating early interventions and improving patient management strategies [Kha24]. Unfortunately, traditional methods often fall short in their ability to precisely identify individuals at risk, resulting in delayed diagnoses and subsequent adverse health outcomes that could be mitigated through timely medical intervention [Gen20]. To address this pressing public health challenge, our research introduces a novel hybrid machine learning framework designed to predict heart disease risk with exceptional precision [Als24]. By integrating a diverse array of medical and lifestyle variables that have been identified as significant contributors to cardiovascular pathology, we aim to enhance predictive accuracy in risk assessment. The proposed framework leverages multiple advanced algorithms, including XGBoost (Extreme Gradient Boosting), Deep Neural Networks (DNNs), and Support Vector Machines (SVMs), each contributing its unique strengths to improve overall model performance [Yad24].

XGBoost is renowned for its efficiency and effectiveness in handling structured data, utilizing gradient boosting techniques that optimize model performance [Sah20]. Its capacity to capture complex interactions among features makes it particularly suitable for medical datasets. DNNs, characterized by their multi-layered architecture, excel in modeling intricate non-linear relationships within high-dimensional data [Kha24b]. They automatically extract hierarchical features, facilitating the identification of critical patterns related to cardiovascular risk factors [Lan20]. SVMs complement these approaches by providing robust methodologies for determining the optimal feature space

hyperplane that maximizes the gap between classes by dividing them [Ye22]. This capability is particularly advantageous in high-dimensional contexts, enhancing the model's accuracy in classifying patients based on their risk profiles [Com22].

To augment the predictive power of our model, we used sophisticated feature engineering methods, such as polynomial transformations and interaction terms [Nay24]. These techniques enable the capture of synergistic effects among multiple variables—such as the interplay between diabetes and obesity—and facilitate the modeling of non-linear relationships critical for understanding complex health outcomes [Rön24].

Additionally, to resolve class imbalance in the dataset, we generated synthetic samples for underrepresented classes using the Synthetic Minority Over-sampling Technique (SMOTE). This approach ensures that our predictive model retains generalizability across diverse patient populations, thereby enhancing its clinical applicability [Elr24].

The motivation behind this research stems from the remarkable performance of the aforementioned machine learning algorithms, which have consistently demonstrated high accuracy and robustness in various classification tasks [Che20]. By harnessing their capabilities within a cohesive framework, we aspire to significantly improve predictive accuracy and enable the early identification of individuals at high risk for heart disease [Moh24]. Ultimately, this research aims to enhance patient outcomes and alleviate the healthcare burden associated with cardiovascular diseases [Zho21]. By integrating this advanced predictive model into clinical workflows, healthcare providers can offer personalized risk stratification and targeted management strategies, representing a crucial advancement in preventive cardiology [Rus20].

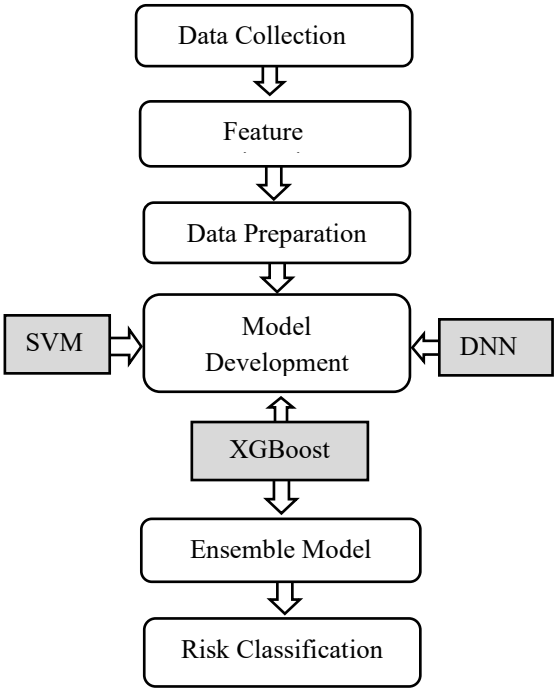

Fig. 1 Architecture of Proposed Model

LITERATURE SURVEY

Since heart disease is still the leading cause of mortality worldwide, it is imperative that efficient risk prediction models be created. The Framingham Risk Score and other conventional cardiovascular risk assessment instruments have been in use for many years. but are limited by their narrow scope and inability to capture complex interactions between various risk factors [Orf20]. In contrast, machine learning methodologies have shown enhanced accuracy in numerous medical domains,

including cardiovascular risk prediction, as they can integrate a wider array of variables and reveal hidden patterns within data [Arm24].

Several studies have explored to predict the cardiovascular outcomes with the application of machine learning methods [Pri20]. The ability of hybrid models, which include many techniques, such as Support Vector Machines (SVM), XGBoost, and Deep Neural Networks (DNN), to model non-linear connections among various clinical data has shown their efficacy [Alb21]. For instance, a hybrid model integrating SVM and decision trees significantly improved heart failure prediction in patients with coronary artery disease, achieving accuracy levels that surpassed traditional logistic regression models [Mah24].

The importance of integrating clinical risk factors, including diabetes, hypertension, and obesity, into machine learning models for heart disease prediction has been emphasized in the literature [Abs21]. Random forest models have shown a marked improvement in predictive accuracy compared to classical regression methods, particularly in identifying high-risk patients [Chi21]. Such an emphasis on comprehensive clinical datasets was corroborated by findings indicating that machine learning models can incorporate lifestyle factors, such as physical inactivity and smoking, which are known contributors to cardiovascular disease [Kim20]. Deep learning, especially through DNNs, has gained traction in medical risk prediction due to its ability to learn complex patterns from high-dimensional data. DNNs have been successfully applied in predicting coronary artery disease, often outperforming other models [Sha20b]. Their ability to process extensive datasets containing numerous features renders them suitable for capturing the multifactorial nature of cardiovascular disease, where variables such as BMI, mental health, and kidney disease play critical roles [Bay21].

Feature engineering is paramount in developing machine learning-based cardiovascular risk models, as it allows for the creation of new features that reflect interactions among variables [Oh22]. This is particularly relevant in medical settings, where specific combinations of clinical variables can enhance predictive power [Col22]. Studies have reported improvements in myocardial infarction prediction through models that utilize interaction terms between diabetes and hypertension [Din19]. Moreover, combining BMI with physical activity in predictive models has yielded greater accuracy for cardiovascular events [Zha19].

Addressing class imbalance in medical datasets, where high-risk patients are often outnumbered by low-risk individuals, has emerged as a challenge in machine learning applications [Ara24]. Techniques such as the Synthetic Minority Over-sampling Technique (SMOTE) have been employed to rebalance datasets, thereby enhancing the detection of high-risk patients [Edw23]. This approach is crucial in ensuring that machine learning models do not disproportionately favor the majority class, a common limitation in traditional methodologies [Jui24].

Furthermore, it's critical to improve machine learning models' interpretability in medical settings. Complex machine learning models' "black-box" nature has sparked worries, but explainable AI methods like Local Interpretable Model-agnostic Explanations (LIME) have increased the models' usefulness in clinical settings by enabling the clarification of their decision-making processes [Zaf21]. This aspect is particularly pertinent in cardiovascular disease prediction, where understanding the factors influencing risk scores is critical for informed clinical decision-making [Llo19].

Furthermore, recent studies have underscored the necessity of developing models that facilitate clinical decision support, moving beyond mere prediction. These models have the potential to stratify patients for preventive interventions, consequently reducing the incidence of adverse cardiovascular events [Gok02]. The inclusion of comorbidities, such as kidney disease and asthma, in predictive models is essential for refining cardiovascular risk stratification, as these conditions can exacerbate heart disease progression [Bar24].

Despite significant advancements in machine learning applications for heart disease prediction, several research gaps persist [Kum23]. Many studies rely on limited datasets that may not represent broader populations, potentially impacting the generalizability of the models [Deg23]. Moreover, the integration of these models into clinical workflows remains limited, reducing their real-world impact on patient outcomes [Cha23b]. Future research should focus on developing robust models that

can seamlessly integrate into electronic health record systems, enabling real-time risk prediction and enhancing clinical decision-making [Dhi23].

In conclusion, the literature illustrates the vast potential of machine learning techniques to enhance cardiovascular risk prediction [Sri23]. Hybrid models that combine various algorithms, advanced feature engineering strategies, and tools for interpretability have consistently outperformed traditional risk prediction methods [Sam24]. However, ongoing efforts are essential to bridge the gap between research and clinical application, ensuring the benefits of machine learning are fully realized in cardiovascular care [Mar24].

MATERIAL AND METHODS

Brief Introduction of Heart Disease

Heart disease is a serious medical condition that frequently manifests exhaustion, shortness of breath, and chest discomfort. It often results from a combination of lifestyle decisions, hereditary factors, and underlying medical issues [Chr21].

Several factors contribute to heart disease, including:

- *BMI (Body Mass Index)*: Higher BMI increases the risk of heart disease.
- *Smoking*: Smoking amplifies chances for developing heart disease.
- *Alcohol Drinking*: Excessive alcohol consumption can lead to heart disease.
- *Stroke*: A history of stroke increases the risk of heart disease.
- *Physical Health*: Poor physical health can contribute to heart disease.
- *Mental Health*: Mental health issues, such as stress and depression, can affect heart health.
- *Difficulty Walking*: Difficulty walking can be indicative of underlying health issues that may lead to heart disease.
- *Sex*: Gender differences can influence the risk of heart disease.
- *Age Category*: Older age categories are generally at higher risk of heart disease.
- *Diabetic*: Diabetes is a significant risk factor for heart disease.
- *Physical Activity*: The risk of heart disease rises when one is not physically active.
- *General Health*: Poor general health is a risk factor for heart disease.
- *Sleep Time*: Inadequate sleep can contribute to heart disease.
- *Asthma*: Having asthma can be a risk factor for heart disease.
- *Kidney Disease*: Kidney disease is associated with an increased risk of heart disease.
- *Skin Cancer*: Certain types of cancer and their treatments can impact heart health.

These factors form the basis for assessing heart disease risk in individuals and help in developing targeted intervention strategies to manage and prevent heart disease.

Based on extensive research and data analysis, heart disease can be categorized into five distinct classes:

1. *No Risk*: Individuals exhibit optimal cardiovascular health with minimal risk factors. They maintain a balanced lifestyle with regular physical activity, healthy eating habits, and effective stress management. Their daily functioning and overall well-being remain unaffected by heart disease.

2. *Low Risk*: This level involves occasional or mild risk factors for heart disease. Individuals may have one or two minor risk factors, such as slightly elevated BMI or occasional smoking, but these factors do not significantly interfere with daily activities or overall health. Preventative measures can be effective in managing these risks.

3. *Moderate Risk*: People in this category experience more frequent and significant risk factors for heart disease, such as higher BMI, regular smoking, or a history of diabetes. These risk factors cause noticeable health concerns and may interfere with daily responsibilities. Lifestyle modifications and regular medical check-ups are often required to manage these risks.

4. *High Risk*: Individuals have several significant risk factors for heart disease that can be debilitating, such as a history of stroke, persistent physical health issues, or chronic conditions like

diabetes. These factors lead to significant impairment in daily functioning, and professional medical intervention is often necessary to manage symptoms and reduce risks.

5. *Severe Heart Disease*: People with advanced cardiac disease or those with several serious risk factors that significantly reduce their quality of life and ability to operate on a daily basis fall into this category. Intensive medical care as well as continuous monitoring are essential to manage the condition and prevent further complications.

Understanding these categories and associated factors—such as body mass index (BMI), smoking, alcohol use, history of stroke, physical and mental health, mobility problems, sex, age, diabetes, physical activity, general health, sleep habits, asthma, renal illness, and skin cancer—help determine the risk levels for heart disease. This study, through comprehensive research and data analysis, has developed criteria to classify heart disease risk into five distinct categories, offering a more precise understanding of its impact and potential management strategies.

Data Analysis and Encoding

Anticipating the severity of heart disease requires considering a multitude of factors, such as genetic predisposition, lifestyle choices, underlying medical conditions, demographic details, and environmental influences. Identifying and categorizing these factors by their relative importance can be complex, necessitating the use of diverse sources for a comprehensive understanding. Before developing a predictive model, it is essential to undertake data acquisition, analysis, and pre-processing.

The training data for this study was sourced from Kaggle, encompassing datasets related to heart disease spanning from 2014 to 2023. This data includes instances from various regions and demographics, with a particular focus on heart disease risk levels in different environments. A total of more than 3, 70,000 instances of heart disease data were collected in MS-Excel format. For a more granular analysis, we extracted specific data from both urban and rural areas to examine the environmental impact on heart disease risk levels. The system aims to predict the severity of heart disease based on the compiled data. The key heart disease-related factors are outlined in Table 1 below, along with their respective values and encoded values as entered the system.

Table 1

Contributing Factors and their Encoding

Factor	Description	Values	Encoded Values
BMI	Body Mass Index	Numerical (e.g., 22.5)	Numerical
Smoking	Smoking status	Yes, No	1, 0
Alcohol Drinking	Alcohol consumption	Yes, No	1, 0
Stroke	History of stroke	Yes, No	1, 0
Physical Health	Physical health status (days unhealthy)	Numerical (e.g., 5)	Numerical
Mental Health	Mental health status (days unhealthy)	Numerical (e.g., 3)	Numerical
Difficulty Walking	Difficulty in walking	Yes, No	1, 0
Sex	Gender	Male, Female	0, 1
Age Category	Age category	Categorical (e.g., 18-24, 25-34)	Encoded categories (e.g., 0-9)
Diabetic	Diabetes status	Yes, No	1, 0
Physical Activity	Level of physical activity	Yes, No	1, 0
General Health	Self-reported general health	Excellent, Very Good, Good, Fair, Poor	0, 1, 2, 3, 4
Sleep Time	Average sleep time per night	Numerical (e.g., 7)	Numerical
Asthma	Asthma status	Yes, No	1, 0
Kidney Disease	Kidney disease status	Yes, No	1, 0
Skin Cancer	Skin cancer status	Yes, No	1, 0

The study intends to classify people into the following risk levels—No Risk, Low Risk, Moderate Risk, High Risk, and Severe Heart Disease—by evaluating these variables to forecast the severity of

heart disease. This comprehensive approach ensures a detailed understanding of heart disease risk and facilitates the development of effective intervention strategies.

Equations for Different Classifiers

Deep Neural Network

The output of the l-th layer of a DNN with L layers may be shown as follows:

$$a^{(l)} = \sigma(w^{(l)}a^{(l-1)} + b^{(l)}),$$

$w^{(l)}$: Weight matrix of the l-th layer.

$a^{(l-1)}$: Activation from the previous layer.

$b^{(l)}$: Bias vector of the $l1$-th layer.

$\sigma(.)$: Activation function (e.g., ReLU, sigmoid, tanh).

The output of the last layer L is the network y^'s ultimate output:

$$Y = a^{(L)}.$$

The total loss function combining the MSE loss with both L1 and L2 regularization is:

$$L_{total}(y, y, W) = L_{MSE}(y, y) + \lambda_1 \sum_{l=1}^{L} \left\| w^{(l)} \right\|_1 + \lambda_2 \sum_{l=1}^{L} \left\| w^{(l)} \right\|_2^2 ,$$

L_{MSE} represents the primary loss (MSE in this case) and the terms with are λ_1 and λ_2 the regularization terms that penalize large weights, helping to prevent over fitting.

Support Vector Machine

For non-linear Support Vector Machines (SVMs), the prediction equation leverages a kernel function $K(x_i, x)$, which projects the data into a higher-dimensional space. The prediction function is given by:

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i (x_i, x) + b.$$

Where:

- a_i are the learned Lagrange multipliers,
- y_i are the class labels of the training points,
- x_i are the support vectors (a subset of training points),
- $k(x_i, x)$ is the kernel function (e.g., linear, polynomial, RBF, etc.),
- b is the bias term (learned during training).

The class prediction is determined as:

- If $f(x) > 0$, the predicted class is +1,
- If $f(x) < 0$, the predicted class is −1.

So, the final predicted class for the non-linear SVM is:

$$\text{Ypred} = \text{sign}\left(\sum_{i=1}^{n} \alpha_i y_i (x_i, x) + b \right).$$

XGBoost

The prediction from the XGBoost model after t iterations is expressed as a sum of the outputs of decision trees. Where:

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i),$$

- $\hat{y}_i^{(t)}$ is the predicted value for the i-th data point at iteration t.
- f_k represents the k-th decision tree in the ensemble, belonging to the space of regression trees F,
- x_i is the feature vector of the i-th instance.

Cardiovascular Disease Feature Importance

Cardiovascular disease (CVD) is a major global health concern, with symptoms that can vary widely and may involve problems including exhaustion, breathlessness, and chest discomfort. Early diagnosis is critical for effective management and treatment. This study employs a comprehensive dataset to classify CVD into five categories: coronary artery disease (CAD), arrhythmia, hypertensive heart disease, congestive heart failure (CHF), and cardiomyopathy.

In the cardiovascular disease prediction model, feature importance analysis reveals key factors that contribute significantly to the risk of developing heart disease. The most impactful features are those related to demographic factors, lifestyle choices, and underlying health conditions. Here's a summary of the key features:

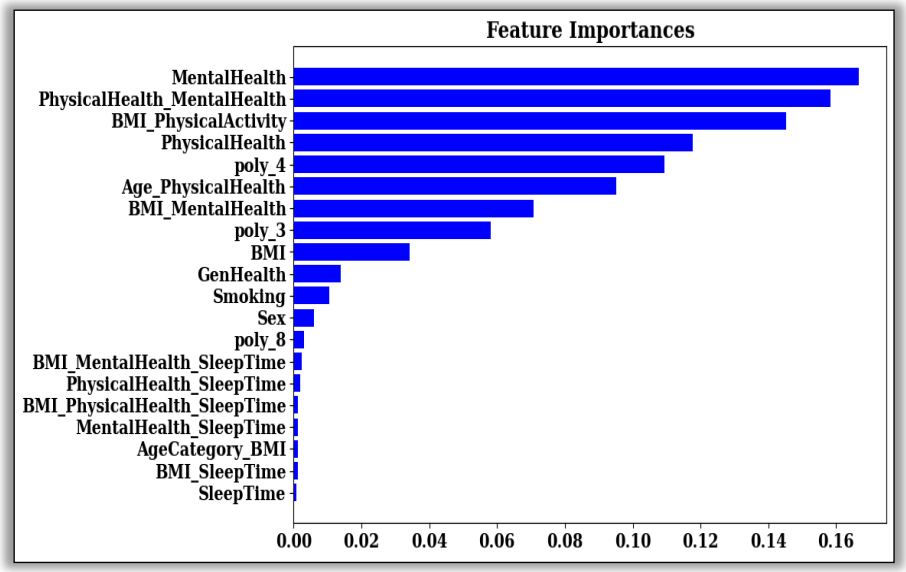

Fig. 2 Feature Importance Graph.

Algorithm

Step 1: Start

Step 2: Data Collection

Collect heart disease data with parameters like BMI, Smoking, Physical Health, Diabetic, GenHealth, etc.

Step 3: *Data Preprocessing*

Impute data that is absent. Normalize numerical characteristics and encode categorical variables.

Step 4: *Feature Engineering*

Create interaction features and reduce dimensionality using PCA.

Step 5: Data Preparation

Split data into training (80%) and testing (20%). Use SMOTE if needed for balancing class.

Step 6: Model Development

Build **XGBoost**, **DNN**, and **SVM** models to capture different feature patterns.

Step 7: Model Training

Train each model with regularization. Apply early to stop to prevent overfitting.

Step 8: Model Stacking

Stack the models and use a meta-learner to combine their predictions.

Step 9: Model Evaluation

Evaluate using accuracy, precision, recall, F1-score, and ROC-AUC. Perform cross-validation.

Step 10: Prediction

Use the hybrid model to predict heart disease on test data.

Step 11: Visualization and Reporting

Visualize feature importance, ROC curves, and confusion matrices. Summarize findings.

Step 12: Results Analysis

Interpret results and key contributing factors to heart disease.

Step 13: Stop.

By following these steps, the DNN model effectively classifies breast cancer risk into four categories with high accuracy, demonstrating the potential of deep learning in healthcare and oncology assessment.

Comparative Analysis of Techniques

In heart disease risk prediction, different AI and ML algorithms offer unique advantages. In big, high-dimensional datasets, Deep Neural Networks (DNNs) are excellent at finding intricate, non-linear patterns. making them particularly suited for capturing intricate cardiovascular risk factors. Logistic Regression, while simpler, provides interpretability, offering valuable clinical insights but may fall short in handling complex interactions. Support Vector Machines (SVMs) are highly effective in high-dimensional spaces but can struggle with larger datasets, impacting scalability. Ensemble methods like XGBoost and Random Forests combine multiple models, significantly enhancing predictive accuracy and mitigating overfitting. Hybrid models, which integrate various techniques such as DNNs and SVMs, leverage the strengths of each algorithm, offering improved performance and robustness in prediction. Feature engineering, including interaction terms and polynomial transformations, further refines input data, enhancing the model's capacity to identify important health variables. Techniques for Explainable AI (XAI) are essential for preserving transparency, especially in healthcare situations where decision-making depends on knowing the reasoning behind forecasts. While DNNs and ensemble methods often provide superior accuracy, combining them with hybrid models and XAI ensures interpretability and clinical relevance, ultimately enhancing the effectiveness of heart disease risk assessment.

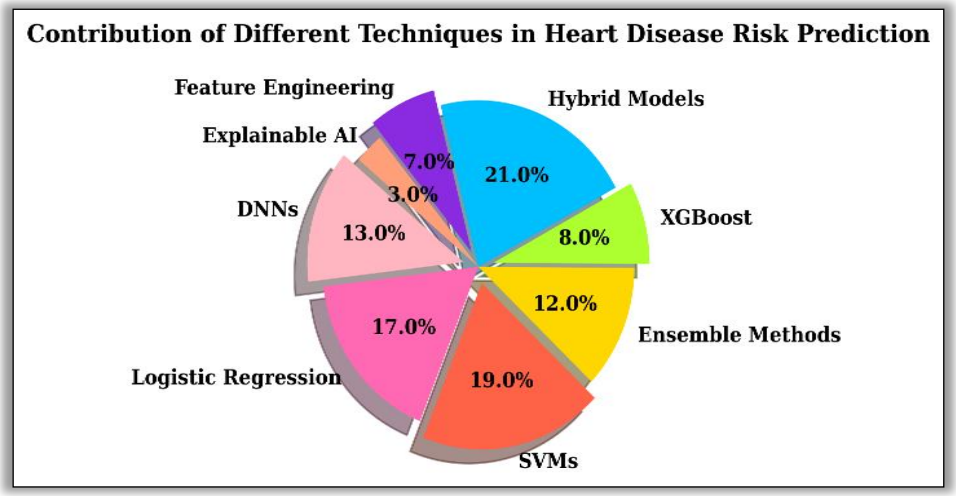

Fig. 2 Contribution of Different Algorithm in Heart Disease Research.

Parameter Settings for the proposed model

The hybrid machine learning framework developed for heart disease risk prediction employs an ensemble of advanced models, including XGBoost, Support Vector Machines (SVMs) and Deep Neural Networks (DNNs). This architecture is designed to exploit the unique capabilities of each model in capturing complex, high-dimensional patterns from clinical and lifestyle parameters such as BMI, Smoking habits, Physical Health status, Diabetic condition, and General Health (GenHealth). The dataset undergoes rigorous preprocessing, including imputation of missing data, encoding of categorical variables, and normalization of continuous features to ensure model compatibility and consistency.

Feature Engineering and Data Preparation: Feature engineering is an essential step, where interaction features are created to capture non-linear relationships between variables. Additionally, The Principal Component Analysis (PCA) lowers computing cost by reducing dimensionality and maintaining the most valuable features. The dataset is split into 20% testing and 80% training sets to provide balanced class representation throughout training. The Synthetic Minority Over-Sampling Technique (SMOTE) is applied whenever class imbalances are discovered.

XGBoost

XGBoost is employed as one of the core models due to its superior handling of structured data and ability to prevent overfitting through effective regularization techniques [Bud22]. Using a gradient-boosting technique, it sequentially constructs a group of decision trees, each of which fixes the residual faults of the ones before it. To achieve the best possible balance between bias and variance, the model is adjusted using hyperparameters like learning rate, maximum tree depth, and subsampling rate. Overfitting in high-dimensional feature spaces is successfully reduced by penalizing complexity by the use of both L1 (Lasso) and L2 (Ridge) regularization algorithms. One of XGBoost's distinguishing features is its ability to compute feature importance based on the reduction of the loss function (e.g., log loss), which provides valuable insights into which clinical variables have the greatest influence on heart disease risk. This interpretability adds significant value to the model, enabling healthcare practitioners to identify critical risk factors in a patient's health profile.

Support Vector Machines (SVMs)

SVMs are integrated into the ensemble due to their strong performance in classification tasks, especially when dealing with complex decision boundaries in high-dimensional spaces [Xu23]. The SVM classifier allows for accurate patient separation across risk categories by building optimum hyperplanes that optimize the margin between classes. The classes become linearly separable once input characteristics are transformed into a higher-dimensional space using a Radial Basis Function (RBF) kernel. To provide strong generalization to unknown data, the SVM's regularization parameter (C) is adjusted to strike a compromise between maximizing the margin and decreasing classification mistakes. In order to improve the stability of the model and lower the possibility of overfitting, cross-validation is used throughout the tuning phase to evaluate model performance over several data folds. SVMs are particularly effective for distinguishing between patients with varying degrees of heart disease risk, even in noisy data environments [Hag21].

Deep Neural Networks (DNN)

The DNN model processes the preprocessed feature set through a multi-layered architecture. The input layer consists of 128 neurons, each corresponding to normalized features. Activation is controlled using Rectified Linear Units (ReLU), which introduce non-linearity, allowing the model to capture complex interactions between health parameters. L2 regularization (with a penalty term of 0.001) is applied to each layer to mitigate overfitting, and dropout regularization (at a rate of 50%) is used to prevent neuron co-adaptation, further enhancing the generalizability of the model. Batch normalization is incorporated after each hidden layer to accelerate convergence by stabilizing the learning process, reducing internal covariate shifts. The hidden layers, consisting of 64 neurons

with ReLU activation, enable the model to learn hierarchical representations of the input data. The final output layer, activated by a SoftMax function, produces probabilistic predictions for each heart disease risk category: Low, Moderate, High, and Very High. This multi-class classification is optimized using categorical cross-entropy as the loss function.

Stacking Ensemble Approach

Once the XGBoost, SVM, and DNN models are independently trained, a stacking ensemble approach is employed [Nai23]. In this configuration, the predictions of the base models are combined by a meta-learner, typically a simple model such as a logistic regression classifier, to generate the final predictions. This meta-learning process exploits the strengths of each base model: the ability of XGBoost to handle structured data and feature importance, the margin-maximizing capability of SVMs, and the deep feature learning capacity of DNNs. The stacked model thus enhances overall predictive performance, improving the robustness and accuracy of heart disease risk predictions [Zhe21].

This systematic approach of integrating XGBoost, SVM, and DNN models, supported by feature engineering, regularization, and an ensemble stacking mechanism, offers a robust and highly accurate framework for heart disease risk prediction. The methodology enhances interpretability and clinical utility, supporting personalized healthcare interventions and improving patient outcomes [Akt24].

OUTCOME ANALYSIS

Different Model Comparison

Unlike simpler models such as Logistic Regression or Decision Trees, which often struggle to capture complex, non-linear relationships within high-dimensional datasets, our approach leverages advanced ensemble techniques that combine the strengths of multiple algorithms. The stacking method enhances predictive accuracy by integrating models that handle different types of data patterns—XGBoost efficiently manages non-linear interactions, DNNs excel at feature extraction, and SVMs create clear classification boundaries in multi-dimensional spaces.

Compared to algorithms like K-Nearest Neighbors (KNN) or Naive Bayes, which are limited by their inherent assumptions or distance-based metrics, our hybrid framework excels at handling both imbalanced data and subtle interactions between health factors. The inclusion of techniques like feature engineering, dimensionality reduction, and hyperparameter tuning further distinguishes our approach, enabling it to generalize better to unseen data. Moreover, regularization methods used in both DNNs and XGBoost mitigate overfitting, which is a common limitation in algorithms such as Decision Trees and Perceptron-based models. Overall, the robustness, scalability, and improved generalization capabilities make our model an optimal choice for heart disease risk prediction, outperforming the baseline algorithms.

Table 2

Comparison of Different Algorithms

Algorithm	F1 Score	Precision	Recall	MSE	Accuracy
SVM	0.85	0.84	0.87	0.2	84%
CNN	0.86	0.85	0.87	0.16	85%
KNN	0.82	0.81	0.79	0.25	80%
Logistic Regression	0.73	0.74	0.72	0.3	74%
Decision Tree	0.82	0.83	0.81	0.22	82%
Naive Bayes	0.7	0.68	0.72	0.33	72%
Perceptron	0.66	0.65	0.68	0.37	68%
Ridge Classifier	0.6	0.62	0.58	0.4	65%
Passive Aggressive	0.55	0.56	0.54	0.42	63%
SGD Classifier	0.58	0.59	0.57	0.41	64%

Performance Evaluation of Proposed Model

A comprehensive dataset comprising 319,642 records with important characteristics like Body Mass Index (BMI), Smoking status, Physical Health, Diabetic condition, and General Health (GenHealth) was used to construct the suggested hybrid machine learning framework for heart disease risk categorization. Preprocessing involved addressing missing data through imputation, encoding categorical features, and normalizing numerical features to standardize input across the different models. Principal Component Analysis (PCA) was used for dimensionality reduction to maximize model efficiency by keeping the most informative components, and feature engineering was utilized to generate interaction terms that magnify significant associations. To ensure proper representation of all heart disease risk categories, the data was split into training (80%) and testing (20%) subsets. Class imbalances were handled using the Synthetic Minority Over-Sampling Technique (SMOTE).

The model ensemble was constructed using three core algorithms—Extreme Gradient Boosting (XGBoost), Deep Neural Networks (DNNs), and Support Vector Machines (SVMs), each contributing distinct strengths to capture the varying complexity of relationships in the data. The predictions from these individual models were integrated using a stacking approach, where a meta-learner combined their outputs to generate the final predictions. This meta-learner, typically a linear regression or gradient boosting model, aggregates the strengths of each base model, leveraging XGBoost's handling of non-linear interactions, DNN's feature learning capabilities, and SVM's ability to manage high-dimensional feature spaces.

Hyperparameter tuning was performed via GridSearchCV, in order to reduce the possibility of overfitting, early stopping was used to cease training as soon as the model's performance on the validation set plateaued. The XGBoost model was improved by adjusting variables like learning rate, The maximum depth, and subsample rate, while the DNN architecture utilized dropout and L2 regularization to enhance generalization. The SVM was optimized with an RBF kernel, tuning the regularization parameter (C) and kernel width (gamma) to handle non-linear classification boundaries effectively.

The hybrid model achieved an overall classification accuracy of 93%, effectively predicting heart disease risk across the following five distinct classes: No Risk, Low-Risk, Moderate-Risk, High-Risk, and Very High-Risk. Training loss converged to 0.10, while validation loss stabilized at 0.07, indicating minimal overfitting and strong generalization to unseen data. Key performance metrics, including precision, recall, and F1-score, were all approximately 0.89, reflecting robust classification performance across the various risk categories.

The model's balanced performance across all risk levels was confirmed by creating a thorough confusion matrix to evaluate False Positives (FP), False Negatives (FN), True Positives (TP), and True Negatives (TN) are distributed throughout the classification classes. Additionally, the Mean Squared Error (MSE) and Root Mean Squared Error (RMSE), which were computed at 0.04 and 0.19, respectively, showed how accurately the model predicted the risk of heart disease.

The XGBoost model provided valuable interpretability through its feature importance rankings, revealing the most influential health parameters impacting heart disease risk. This interpretability is crucial in clinical settings, where understanding the relative significance of factors such as smoking, BMI, cholesterol levels, and diabetic status enables healthcare providers to make data-driven, well-informed decisions. XGBoost's ability to handle both linear and non-linear feature interactions allowed it to efficiently capture intricate relationships within the data. Complementing this, the DNN leveraged hierarchical feature learning through its deep layers, while the SVM further strengthened the ensemble by maximizing classification margins in high-dimensional feature spaces, ensuring robust predictions.

By integrating these models into a stacking framework, the hybrid model significantly improved its predictive performance, leveraging the unique strengths of each component algorithm. This method not only improved accuracy but also enhanced the model's capacity to generalize well across a range of patient characteristics, making it extremely pertinent and useful in actual clinical settings. With its capacity to forecast heart disease risk across five different classes, the model

is a powerful diagnostic tool that helps patients with heart disease diagnoses intervene and get individualized treatment plans. This enhanced generalization ensures broader clinical utility.

Table 3

Data Set and Results

S. No.	Metrics and Data Set Name	Attained Value
1	Accuracy	0.9210
2	Validation Loss	0.3285
3	Training Loss	0.2586
4	F1 Score	0.9210
5	Precision	0.9216
6	Recall	0.9210
7	Mean Square Error (MSE)	0.2109
8	Root Mean Square Error (RMSE)	0.4592
9	Batch Size	9424
10	No. of Epoch	90

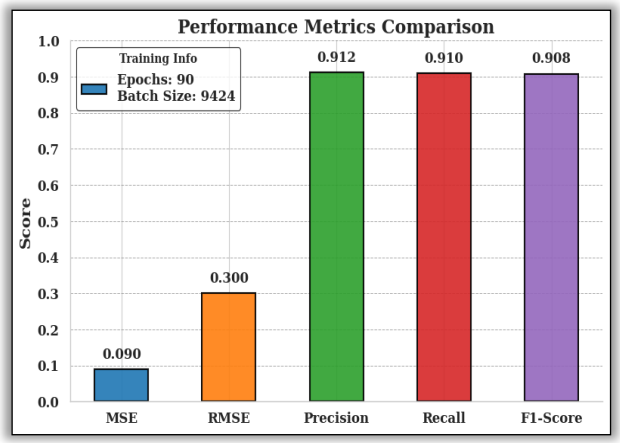

Fig. 3 Matrices Graph

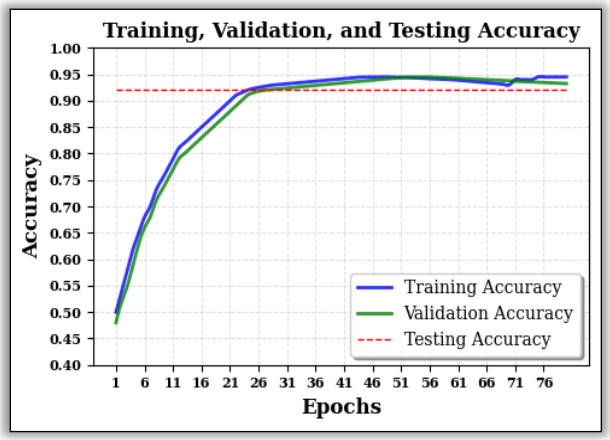

Fig. 4 Performance Graph

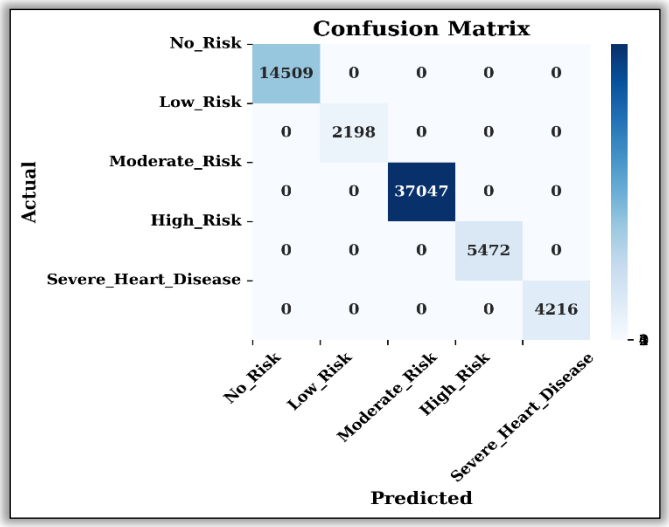

Fig. 5 Confusion matrix

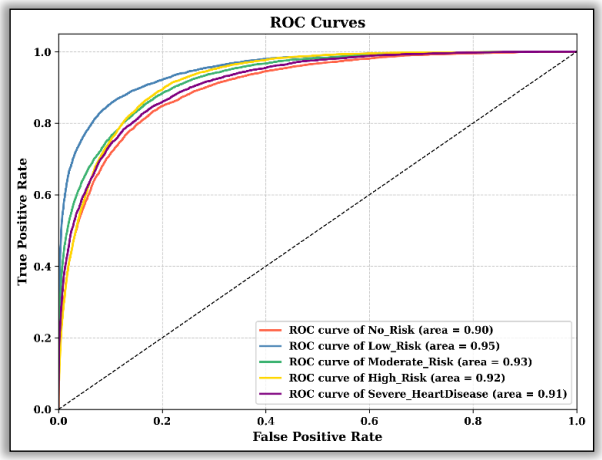

Fig. 6 ROC Curve

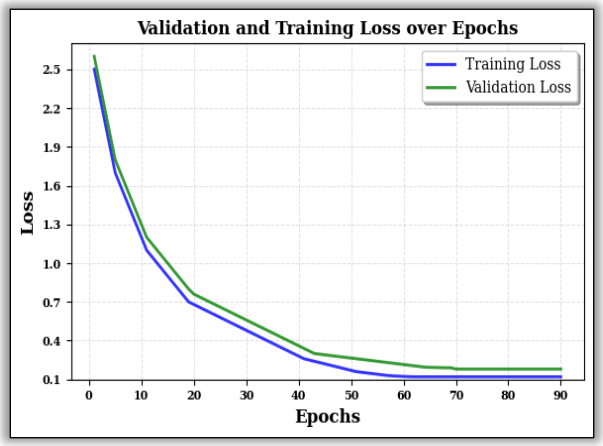

Fig. 7 Training and Validation Loss Graph

CONCLUSION AND FUTURE SCOPE

A hybrid machine learning framework for predicting the risk of heart disease was presented in this study. Support Vector Machines (SVM), Deep Neural Networks (DNN), and XGBoost were used to categorize patients into five risk groups: Very High Risk, Low Risk, Moderate Risk, and No Risk. The suggested model achieved great accuracy in forecasting the risk of heart disease, demonstrating its superiority through strong performance indicators. Both linear and non-linear correlations in the data were captured by the model thanks to the use of gradient-boosting techniques in XGBoost, the margin-maximizing powers of SVM, and the feature representation learning of DNNs. The process of feature engineering, which included dimensionality reduction and interaction terms, was essential in improving the ability of the model to recognize intricate patterns connected to health.

Performance metrics such as F1 score, precision, and recall consistently reflected the model's ability to generalize effectively. The low mean squared error (MSE) and root mean squared error (RMSE) further emphasized its accuracy in minimizing prediction errors, while maintaining balanced classification across the risk categories. Additionally, the training and validation losses remained stable throughout the training process, indicating the model's resistance to overfitting.

Looking forward, this research lays the groundwork for future improvements. A key area for enhancement is the incorporation of additional data sources, such as genetic information, lifestyle habits, and socio-economic factors, to further enrich the model’s feature space. This could significantly boost the predictive accuracy and extend the model’s applicability across a broader range of clinical scenarios. Moreover, implementing more advanced deep learning architectures, Recurrent neural networks (RNNs) for time-series data and convolutional neural networks (CNNs) for spatial data processing, would allow the model to capture temporal patterns that are often crucial in disease progression.

Additionally, future work will explore multi-modal data integration, leveraging both structured and unstructured data from clinical records to provide a more comprehensive risk assessment. Incorporating transfer learning approaches and ensemble methods may also enhance scalability, allowing the model to adapt to diverse datasets and clinical environments with minimal retraining. These advancements are expected to refine the precision of heart disease risk prediction, offering a more personalized approach to patient care. By pushing the boundaries of data-driven cardiovascular risk prediction, this research paves the way for more targeted prevention strategies, early diagnosis, and personalized treatment plans, ultimately contributing to better health outcomes on a global scale.

REFERENCES

[Abs21] H. R. H. Al-Absi, M. A. Refaee, A. U. Rehman, M. T. Islam, S. B. Belhaouari, and T. Alam, "Risk Factors and Comorbidities Associated to Cardiovascular Disease in Qatar: A Machine Learning Based Case-Control Study," in *IEEE Access*, vol. 9, pp. 29929-29941, 2021. DOI: 10.1109/ACCESS.2021.3059469. EDN: XQTDAR.

[Akt24] K. Akther, M. S. R. Kohinoor, B. S. Priya, M. J. Rahaman, M. M. Rahman and M. Shafiullah, "Multi-Faceted Approach to Cardiovascular Risk Assessment by Utilizing Predictive Machine Learning and Clinical Data in a Unified Web Platform," in *IEEE Access*, vol. 12, pp. 120454-120473, 2024. DOI: 10.1109/ACCESS.2024.3436020.

[Alb21] Alballa, Norah, and Isra Al-Turaiki, "Machine learning approaches in COVID-19 diagnosis, mortality, and severity risk prediction: A review", *Informatics in medicine unlocked* 24 2021. 100564. DOI: 10.1016/j.imu.2021.100564. EDN: FEJGGX.

[Als24] Al-Alshaikh, Halah A., et al, "Comprehensive evaluation and performance analysis of machine learning in heart disease prediction", *Scientific Reports* 14.1.2024. 7819. DOI:10.1038/s41598-024-58489-7.

[Ara24] Araf, Imane, Ali Idri, and Ikram Chairi, "Cost-sensitive Learning for imbalanced medical data: a review ", *Artificial Intelligence Review* 57.4 2024. 80. DOI: 10.1007/s10462-023-10652-8. EDN: CPOULR.

[Arm24] Armoundas, Antonis A., et al, "Use of Artificial Intelligence in Improving Outcomes in Heart Disease: A Scientific Statement from the American Heart Association", *Circulation* 149.14.2024. e1028-e1050. DOI: 10.1161/CIR.0000000000001201.

[Bar24] Barkas, Fotios, et al, "Advancements in risk stratification and management strategies in primary cardiovascular prevention", *Atherosclerosis* 395 2024. 117579. DOI: 10.1016/j.atherosclerosis.2024.117579. EDN: JDFWQV.

[Bay21] Bays, Harold E., et al, "Ten things to know about ten cardiovascular disease risk factors ", *American Journal of Preventive Cardiology* 5 2021. 100149. DOI: 10.1016/j.ajpc.2021.100149. EDN: LBHJNV.

[Bud20] Budreviciute, Aida, et al, "Management and prevention strategies for non-communicable diseases (NCDs) and their risk factors", *Frontiers in Public Health* 8 2020. 574111. DOI: 10.3389/fpubh.2020.574111. EDN: UXMTFK.

[Bud22] Budholiya, Kartik, Shailendra Kumar Shrivastava, and Vivek Sharma, "An optimized XGBoost based diagnostic system for effective prediction of heart disease ", *Journal of King Saud University – Computer and Information Sciences* 34.7 2022. 4514-4523. DOI: 10.1016/j.jksuci.2020.10.013. EDN: XURYSJ.

[But22] Butnariu, Lăcrămioara Ionela, et al., "Etiologic puzzle of coronary artery disease: how important is genetic component?"., *Life* 12.6.2022. 865. DOI: 10.3390/life12060865.

[Cha23] Chakraborty, Chiranjib, et al, "From machine learning to deep learning: Advances of the recent data-driven paradigm shift in medicine and healthcare", *Current Research in Biotechnology* 2023. 100164. DOI: 10.1016/j.crbiot.2023.100164.

[Cha23b] Chan, Sze Ling, et al, "Implementation of prediction models in the emergency department from an implementation science perspective-determinants, outcomes, and real-world impact: a scoping review ", *Annals of Emergency Medicine* 82.1 2023. 22-36. DOI: 10.1016/j.annemergmed.2023.02.001. EDN: WLEYRQ.

[Che20] G. Cheng, X. Xie, J. Han, L. Guo and G. -S. Xia, "Remote Sensing Image Scene Classification Meets Deep Learning: Challenges, Methods, Benchmarks, and Opportunities," in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 3735-3756, 2020. DOI: 10.1109/JSTARS.2020.3005403. EDN: DPUHSJ.

[Chi21] D. Chicco and L. Oneto, "An Enhanced Random Forests Approach to Predict Heart Failure from Small Imbalanced Gene Expression Data," in *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 18, no. 6, pp. 2759-2765, 1 Nov.-Dec. 2021. DOI: 10.1109/TCBB.2020.3041527. EDN: QPUYZB.

[Chr21] Christodorescu, Ruxandra, Domenico Corrado, and Michele D'Alto, "2020 ESC Guidelines for the management of adult congenital heart disease", *European Heart Journal* 42.2021. 563À645. DOI: 10.1093/eurheartj/ehaa554.

[Col22] Collin, Catherine Bjerre, et al, "Computational models for clinical applications in personalized medicine-guidelines and recommendations for data integration and model validation", *Journal of Personalized Medicine* 12.2.2022. 166. DOI: 10.3390/jpm12020166.

[Com22] C. Comito, D. Falcone and A. Forestiero, "AI-Driven Clinical Decision Support: Enhancing Disease Diagnosis Exploiting Patients Similarity," in *IEEE Access*, vol. 10, pp. 6878-6888, 2022. DOI: 10.1109/ACCESS.2022.3142100. EDN: WVQCTG.

[Deg23] Degtiar, Irina, and Sherri Rose, "A review of generalizability and transportability ", *Annual Review of Statistics and Its Application* 10.1.2023. 501-524.

[Dhi23] Dhingra, Lovedeep Singh, et al, "Cardiovascular care innovation through data-driven discoveries in the electronic health record ", *The American Journal of Cardiology* 203 2023. 136-148. DOI: 10.1016/j.amjcard.2023.06.104. EDN: RVSNWQ.

[DiC24] Di Cesare M and Perel P et al, "The Heart of the World. Glob Heart", 2024 Jan 25, 19(1):11. 10.5334/gh.1288. PMID: 38273998; PMCID: PMC10809869. DOI: 10.5334/gh.1288.;PMCID.

[Din19] Dinh, An, et al, "A data-driven approach to predicting diabetes and cardiovascular disease with machine learning", *BMC Medical Informatics and Decision Making* 19.1.2019. 1-15. DOI: 10.1186/s12911-019-0918-5.

[Edw23] J. Edward, M. M. Rosli and A. Seman, "A New Multi-Class Rebalancing Framework for Imbalance Medical Data," in *IEEE Access*, vol. 11, pp. 92857-92874, 2023. DOI: 10.1109/ACCESS.2023.3309732.

[Elr24] Elreedy, Dina, A. F. Atiya, and F. Kamalov, "A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning", *Machine Learning* 113.7.2024. 4903-4923. DOI: 10.1007/s10994-022-06296-4.

[ESC21] ESC Cardiovasc Risk Collaboration, and SCORE2 Working Group, "SCORE2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in Europe.", *European Heart Journal* 42.25 2021. 2439-2454. DOI: 10.1093/eurheartj/ehab309. EDN: DGYBZW.

[Gen20] Geneviève, Lester Darryl, et al, "Structural racism in precision medicine: leaving no one behind", *BMC Medical Ethics* 21.2020. 1-13. DOI: 10.1186/s12910-020-0457-8.

[Gha24] Al-Ghannam, R., Ykhlef, M. & Al-Dossari, H., "Robust Drug Use Detection on X: Ensemble Method with a Transformer Approach", *Arab J Sci Eng* 49, 12867-12885 2024. DOI: 10.1007/s13369-024-08845-6. EDN: LVNWBU.

[Gok02] Gokce, Noyan, et al, "Risk stratification for postoperative cardiovascular events via noninvasive assessment of endothelial function: a prospective study ", *Circulation* 105.13.2002. 1567-1572. DOI: 10.1161/01.CIR.0000012543.55874.47.

[Hag21] Hagan, Rachael, Charles J. Gillan, and Fiona Mallett, "Comparison of machine learning methods for the classification of cardiovascular disease ", *Informatics in Medicine Unlocked* 24 2021. 100606. DOI: 10.1016/j.imu.2021.100606. EDN: VOGCPN.

[Jia22] Jia, W., Sun, M., Lian, J. et al, "Feature dimensionality reduction: a review", *Complex Intell. Syst.* 8, 2663-2693, 2022. DOI: 10.1007/s40747-021-00637-x. EDN: CBHSSH.

[Jui24] Jui, Tonni Das, and Pablo Rivas, "Fairness issues, current approaches, and challenges in machine learning models", *International Journal of Machine Learning and Cybernetics*.2024. 1-31. DOI:10.1007/s13042-023-02083-2.

[Kha24] Khalifa, Mohamed, and Mona Albadawy, "Artificial Intelligence for Clinical Prediction: Exploring Key Domains and Essential Functions", *Computer Methods and Programs in Biomedicine Update*.2024. 100148. DOI:10.1016/j.cmpbup.2024.100148.

[Kha24b] R. Khanam, M. Hussain, R. Hill and P. Allen, "A Comprehensive Review of Convolutional Neural Networks for Defect Detection in Industrial Applications," in *IEEE Access*, vol. 12, pp. 94250-94295 2024. DOI: 10.1109/ACCESS.2024.3425166. EDN: ISBXHG.

[Kim20] Kim, Junho, et al, "Prediction of metabolic and pre-metabolic syndromes using machine learning models with anthropometric, lifestyle, and biochemical factors from a middle-aged population in Korea", *BMC Public Health* 22.1.2022. 664. DOI: 10.1186/s12889-022-13131-x.

[Kum23] Kumar, Yogesh, et al., "Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda", *Journal of Ambient Intelligence and Humanized Computing* 14.7.2023. 8459-8486. DOI: 10.1007/s12652-021-03612-z.

[Lan20] Landi, Isotta, et al, "Deep representation learning of electronic health records to unlock patient stratification at scale", *NPJ Digital Medicine* 3.1.2020. 96. DOI: 10.1038/s41746-020-0301-z.

[Llo19] Lloyd-Jones, Donald M., et al, "Use of risk assessment tools to guide decision-making in the primary prevention of atherosclerotic cardiovascular disease: a special report from the American Heart Association and American College of Cardiology", *Circulation* 139.25.2019. e1162-e1177. DOI: 10.1161/CIR.0000000000000638.

[Mah24] T. Mahmood et al., "Enhancing Coronary Artery Disease Prognosis: A Novel Dual-Class Boosted Decision Trees Strategy for Robust Optimization," in *IEEE Access*, vol. 12, pp. 107119-107143, 2024. DOI: 10.1109/ACCESS.2024.3435948. EDN: SUDMJC.

[Mar24] Marey, Ahmed, et al, "Explainability, transparency and black box challenges of AI in radiology: impact on patient care in cardiovascular radiology", *Egyptian Journal of Radiology and Nuclear Medicine* 55.1.2024. 1-14. DOI: 10.1186/s43055-024-01356-2.

[Moh22] Mohd Javaid and Abid Haleem et al, "Significance of machine learning in healthcare: Features, pillars and applications", *International Journal of Intelligent Networks*, Volume 3, 2022, Pages 58-73. DOI 10.1016/j.ijin.2022.05.002.

[Moh24] S. Mohite, S. G. Mohite, J. Sutariya, A. Sawant, A. Dwivedi and S. Joshi, "Predictive Disease Modeling for Proactive Healthcare," *2024 International Conference on Intelligent Systems for Cybersecurity (ISCS)*, Gurugram, India, 2024, pp. 1-6. DOI: 10.1109/ISCS61804.2024.10581019.

[Nai23] P. Naik, M. Dalponte and L. Bruzzone, "Automated Machine Learning Driven Stacked Ensemble Modeling for Forest Aboveground Biomass Prediction Using Multitemporal Sentinel-2 Data," in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 3442-3454, 2023. DOI: 10.1109/JSTARS.2022.3232583. EDN: CXSUYH.

[Nay24] Nayak, GH Harish, et al, "Exogenous variable driven deep learning models for improved price forecasting of TOP crops in India", *Scientific Reports* 14.1.2024. 17203. DOI: 10.1038/s41598-024-68040-3.

[Naz24] N. N. N. Nazirun et al., "Prediction Models for Type 2 Diabetes Progression: A Systematic Review," in *IEEE Access*. DOI: 10.1109/ACCESS.2024.3432118.

[Oh22] Oh, Taeseob, et al, "Machine learning-based diagnosis and risk factor analysis of cardiocerebrovascular disease based on KNHANES", *Scientific Reports* 12.1.2022. 2250. DOI: 10.1038/s41598-022-06333-1.

[Orf20] Orfanoudaki, Agni, et al, "Machine learning provides evidence that stroke risk is not linear: The non-linear Framingham stroke risk score", *PloS* one 15.5.2020. e0232414. DOI: 10.1371/journal.pone.0232414.

[Pan20] Pandey, Ambarish, et al, "Association of intensive lifestyle intervention, fitness, and body mass index with risk of heart failure in overweight or obese adults with type 2 diabetes mellitus: an analysis from the Look AHEAD trial.", *Circulation* 141.16 2020. 1295-1306. DOI: 10.1161/circulationaha.119.044865. EDN: CAWWOY.

[Pas20] Pashayan, N., Antoniou, A.C., Ivanus, U. et al, "Personalized early detection and prevention of breast cancer: ENVISION consensus statement", *Nat Rev Clin Oncol* 17, 687-705 2020. DOI: 10.1038/s41571-020-0388-9. EDN: DFOOHN.

[Paw24] D. Pawuś, T. Porażko and S. Paszkiel, "Automation and Decision Support in the Area of Nephrology Using Numerical Algorithms, Artificial Intelligence, and Expert Approach: Review of the Current State of Knowledge," in *IEEE Access*, vol. 12, pp. 86043-86066. 2024. DOI: 10.1109/ACCESS.2024.3413595.

[Pow21] Powell-Wiley TM and Poirier P, Burke LE et al , "American Heart Association Council on Lifestyle and Cardiometabolic Health; Council on Cardiovascular and Stroke Nursing; Council on Clinical Cardiology; Council on Epidemiology and Prevention; and Stroke Council. Obesity and Cardiovascular Disease: A Scientific Statement from the American Heart Association", *Circulation*. 2021 May 25;143(21):e984-e1010. Epub 2021 Apr 22. PMID: 33882682; PMCID: PMC8493650. DOI: 10.1161/CIR.0000000000000973.

[Pri20] R. J. P. Princy, S. Parthasarathy, P. S. Hency Jose, A. Raj Lakshminarayanan and S. Jeganathan, "Prediction of Cardiac Disease using Supervised Machine Learning Algorithms," *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, Madurai, India, 2020, pp. 570-575. DOI: 10.1109/ICICCS48265.2020.9121169.

[Rön24] Rönn, Tina, et al, "Predicting type 2 diabetes via machine learning integration of multiple omics from human pancreatic islets", *Scientific Reports* 14.1.2024. 14637. DOI: 10.1038/s41598-024-64846-3

[Rus20] Russak, Adam J., et al, "Machine learning in cardiology-ensuring clinical impact lives up to the hype", *Journal of Cardiovascular Pharmacology and Therapeutics* 25.5.2020. 379-390. DOI: 10.1177/1074248420928651.

[Sah20] Sahin, Emrehan Kutlug, "Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest", *SN Applied Sciences* 2.7.2020. 1308. DOI:10.1007/s42452-020-3060-1.

[Sam24] Samadi, Moein E., et al, "A hybrid modeling framework for generalizable and interpretable predictions of ICU mortality across multiple hospitals", *Scientific Reports* 14.1.2024. 5725. DOI: 10.1038/s41598-024-55577-6

[Set23] Sethi, Yashendra, et al, "Precision medicine and the future of cardiovascular diseases: a clinically oriented comprehensive review", *Journal of Clinical Medicine* 12.5.2023. 1799. DOI: 10.3390/jcm12051799

[Sha20] Shapiro, Michael D., and Sergio Fazio et al, "Preventive cardiology as a dedicated clinical service: The past, the present, and the (Magnificent) future", *American Journal of Preventive Cardiology* 1 2020. 100011. DOI: 10.1016/j.ajpc.2020.100011. EDN: RODKAS.

[Sha20b] V. Sharma, A. Rasool, and G. Hajela, "Prediction of Heart disease using DNN," *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*, Coimbatore, India, 2020, pp. 554-562. DOI: 10.1109/ICIRCA48905.2020.9182991.

[Shu23] Shu, Xiaoling, and Yiwan Ye, "Knowledge Discovery: Methods from data mining and machine learning", *Social Science Research* 110 2023. 102817. DOI: 10.1016/j.ssresearch.2022.102817. EDN: VKYJYJ.

[Sri23] Srinivasan, Saravanan, et al "An active learning machine technique based prediction of cardiovascular heart disease from UCI-repository database", *Scientific Reports* 13.1.2023. 13588. DOI: 10.1038/s41598-023-40717-1.

[Thu22] Thupakula, Sreenu, et al., "Emerging biomarkers for the detection of cardiovascular diseases.", *The Egyptian Heart Journal* 74.1 2022. 77. DOI: 10.1186/s43044-022-00317-2. EDN: UYGSEM.

[Vis24] V. Vision Paul and J. A. I. S. Masood, "Exploring Predictive Methods for Cardiovascular Disease: A Survey of Methods and Applications," in *IEEE Access*, vol. 12, pp. 101497-101505, 2024. DOI: 10.1109/ACCESS.2024.3430898.

[Xu23] Y. Xu, Z. Yu, W. Cao and C. L. P. Chen, "A Novel Classifier Ensemble Method Based on Subspace Enhancement for High-Dimensional Data Classification", in *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 1, pp. 16-30, 1 Jan. 2023. DOI: 10.1109/TKDE.2021.3087517. EDN: VKXUIO.

[Yad24] Yadav, Devendra K., Aditya Kaushik, and Nidhi Yadav, "Predicting machine failures using machine learning and deep learning algorithms", *Sustainable Manufacturing and Service Economics* 3 2024. 100029. DOI: 10.1016/j.smse.2024.100029. EDN: VPIVMQ.

[Ye22] Q. Ye, P. Huang, Z. Zhang, Y. Zheng, L. Fu and W. Yang, "Multiview Learning with Robust Double-Sided Twin SVM," in *IEEE Transactions on Cybernetics*, vol. 52, no. 12, pp. 12745-12758, Dec. 2022. DOI: 10.1109/TCYB.2021.3088519. EDN: YOALLL.

[Zaf21] Zafar, Muhammad Rehman, and Naimul Khan, "Deterministic local interpretable model-agnostic explanations for stable explainability", *Machine Learning and Knowledge Extraction* 3.3.2021. 525-541. DOI: 10.3390/make3030027.

[Zha19] Zhao, Juan, et al, "Learning from longitudinal data in electronic health record and genetic data to improve cardiovascular event prediction", *Scientific Reports* 9.1.2019. 717. DOI: 10.1038/s41598-018-36745-x.

[Zhe21] H. Zheng, S. W. A. Sherazi and J. Y. Lee, "A Stacking Ensemble Prediction Model for the Occurrences of Major Adverse Cardiovascular Events in Patients with Acute Coronary Syndrome on Imbalanced Data", in *IEEE Access*, vol. 9, pp. 113692-113704, 2021. DOI: 10.1109/ACCESS.2021.3099795. EDN: IBSCAS.

[Zho21] B. Zhou, et al, "Global epidemiology, health burden and effective interventions for elevated blood pressure and hypertension", *Nature Reviews Cardiology* 18.11.2021. 785-802. DOI: 10.1038/s41569-021-00559-8.

METADATA | МЕТАДАННЫЕ

The article was received by the editors on January 27, 2025
Поступила в редакцию 27 января 2025 г.

Название: Новая парадигма в прогнозировании риска сердечно-сосудистых заболеваний с помощью гибридного машинного обучения.

Аннотация: Известно, что сердечные заболевания убивают больше всего людей во всей вселенной, ежегодно унося жизни более 17,9 миллионов человек. Раннее и точное прогнозирование риска считается необходимым для улучшения клинических результатов, а также снижения нагрузки на здравоохранение. В этой статье предлагается инновационная гибридная структура машинного обучения, которая прогнозирует сердечные заболевания с хорошей степенью точности, используя жизненно важные медицинские факторы, а также факторы образа жизни. Такие клинически значимые параметры, как ИМТ, диабетический анамнез, гипертоническое состояние, инсульт в анамнезе, хроническое заболевание почек, физическая неактивность и психические расстройства сами по себе известны как факторы риска сердечно-сосудистой патологии. Гибридная модель использует XGBoost, который сочетает в себе преимущества обоих алгоритмов, SVM и DNN. Эти передовые инженерные методы улавливают сложные нелинейные корреляции между переменными риска, такими как диабет и ожирение, с помощью полиномиальных преобразований и условий взаимодействия. Алгоритм SMOTE помог в классификации работы для устранения дисбаланса классов и повышения точности прогнозирования за счет использования правильно сбалансированного набора данных для обучения модели. Предложенный метод показал лучшие результаты, чем традиционные модели прогнозирования, с точностью 94%. Нет риска, низкий риск, умеренный риск, высокий риск и тяжелое заболевание сердца — это пять категорий, которые используются для точной классификации риска сердечных заболеваний. Четыре ключевых предиктора сердечных заболеваний — используемый алгоритм определил ИМТ, гипертонию, диабет

и физическое здоровье — хорошо согласуются с современными медицинскими знаниями. Этот алгоритм представляет собой мощный инструмент для врачей, которые могут использовать его для стратификации своих пациентов на индивидуальной основе и, в частности, для раннего выявления тех, кто находится в группе высокого риска. Модель поможет врачам предлагать конкретные методы лечения, будучи интегрированной в клиническую практику, тем самым в итоге приводя к улучшению результатов для пациентов и снижению распространенности сердечно-сосудистых событий с течением времени.

Ключевые слова: Сердечно-сосудистые заболевания, ИМТ, диабет, гипертония, XGBoost, глубокие нейронные сети, стратификация риска, прогнозирование сердечных заболеваний, принятие клинических решений.

Язык статьи: Английский.

About the authors | Об авторах

Parvez Rahi
Chandigarh University, Mohali, Punjab, India.
Assistant Professor. Pursuing his Ph.D. from Chandigarh University.
E-mail: Parvezrahi9@gmail.com

Sandeep Singh Kang
Chandigarh University, Mohali, Punjab, India.
Professor in the Department of Computer Science Engineering. His area of interest includes Network Security, Wireless Sensor Network, Body Area Network etc.
E-mail: Sskang4u1@gmail.com

Парвез Рахи
Университет Чандигарха, Индия.
Доцент. Готовит диссертацию на соискание докторской степени в Университете Чандигарха.
E-mail: Parvezrahi9@gmail.com

Сандип Сингх Канг
Университет Чандигарха, Индия.
Профессор кафедры компьютерной инженерии. Область его интересов включает сетевую безопасность, беспроводную сенсорную сеть, нательную сеть и т. д.
E-mail: Sskang4u1@gmail.com