Minimax Theory and its Applications
Volume 11 (2026), No. 2s, 200 to 220

Adaptive Sequence Pattern Based Hyperfeature Evaluation Framework for Heart Disease Risk Prediction in Diabetic Patients

K.Sathiesh Kumar[1], Dr.Geetha. K[2]

[1]Research Scholar, Department of Computer Science, Sri Krishna Adithya College of Arts and Science, Coimbatore, Email: k.sathieshr@gmail.com, ORCID: 0009-0005-3235-0392
[2]Associate Professor, Department of Computer Applications, Sri Krishna Adithya College of Arts and Science, Coimbatore, Email: kkgeetha17@gmail.com, Orcid ID: 0000-0003-4638-1235

Abstract
Heart disease is considered to be one of the main causes of death among diabetic patients and, therefore, early identification and evaluation of risks are essential in case of successful clinical intervention. The research proposes a novel structure of heart disease prediction through the datasets related to diabetes through the combination of the advanced preprocessing and feature selection techniques. First, an Adaptive Sequence Pattern Preprocessor (A-SPP) is used to improve the quality of data by means of deleting duplicate records, managing missing values, outliers, and transforming categorical attributes into numerical ones. Moreover, normalization is used to have equal-valued features and to enhance learning of models. The HyperFeature Evaluation Framework (HFEF) is then presented following preprocessing where the most informative and stable clinical attributes are determined on the basis of predictive contribution, stability, and redundancy analysis. Such a strategy helps to decrease the data dimensions and maintain medically significant features. Empirical testing of the framework of two benchmark datasets show that the proposed framework has a significant positive impact on the prediction performance with the accuracy scores of 0.939 and 0.943, respectively. In general, the proposed framework offers an effective and stable solution to the risk prediction of early heart disease in diabetic patients, which could contribute to better clinical practices.

*Keywords:*Diabetes, Feature Selection, Heart Disease Prediction, Machine Learning, Preprocessing

1. Introduction

Heart disease is a leading cause of death in any part of the world and prevalence in the case of diabetic patients is very high. Comorbidity of diabetes and cardiovascular complications has been found to expose diabetics to more risks of heart failure, coronary artery disease and other profound health complications so early prediction is crucial in clinical management [1]. Researchers have found that a number of physiological and biochemical factors including glucose level, insulin resistance, and inflammatory biomarkers are an important determinant of cardiovascular risk in diabetic patients [2], [3]. Thus, proper screening of the risk of heart disease in diabetics has become a significant research issue in contemporary health care systems.

The recent progress in terms of artificial intelligence and Machine Learning (ML) has facilitated the creation of predictive information on determining cardiovascular risk based on medical data. It has also been shown that AI-based prediction models will be promising to detect coronary heart disease in patients with type 2 diabetes by examining both clinical and physiological features [4]. Moreover, several ML methods have been used in the healthcare application of diabetes diagnosis and cardiovascular risk prediction [5-7]. A number of works have also addressed the area of clinical decision support systems and Deep Learning (DL) to enhance the precision of heart disease diagnosis [8]. The systematic reviews also suggest ML methods can substantially improve prediction of cardiovascular risks related to diabetes through the detection of complex associations among clinical variables [9-11].

Even with these developments, most of the current strategies continue to have a number of limitations. Conventional predictive models can be characterized as being based on traditional preprocessing techniques that might not be suitable to deal with noisy data, missing values, and redundant attributes found in medical data. Moreover, most of the attribute selection methods do not assess the stability and clinical significance of attributes, which may impact on prediction accuracy and reliability of the model adversely [12]. More complicated architectures like blockchain-based healthcare systems or fog computing environment are also

introduced by some frameworks and can raise computational overhead and restrict their application to real clinical environments [13].

The other problem is to determine the most informative features which would lead to prediction of heart diseases among patients with diabetes. Clinical research has shown that cardiovascular risk is affected by a number of interconnected factors and selecting features poorly would cause loss of valuable medical data or irrelevant attributes to be included [14-16]. Consequently, the knowledge gap in building an adaptive structure, which combines efficient preprocessing and efficient feature assessment methods to suit diabetes-related heart disease prediction, is present.

To overcome these limitations, this research proposes a novel framework which is a combination of the Adaptive Sequence Pattern Preprocessor (A-SPP) and the HyperFeature Evaluation Framework (HFEF). The A-SPP module improves the quality of data through eliminating redundant records, processing the missing data, outliers, categorical attribute encoding, and normalization of feature values. The HFEF module then features stability based feature assessment to determine the top informative and clinically significant attributes to be used in prediction. The main value of this research is that it will combine a clinically based preprocessing system with an advanced Feature Selection (FS) system to enhance reliability and accuracy of heart disease prediction in diabetics.

The rationale of this research is to come up with a strong and interpretable prediction model capable of guiding medical practitioners to identify cardiovascular risk in diabetic patients early enough.

Organization: Section 2 covers the relevant literature on the models of predicting diabetes and heart diseases. Section 3 introduces the suggested methodology, which will involve the A-SPP preprocessing and HFEF FS algorithms. Section 4 defines the results of the experiment and performance assessment. Section 5 is the final section of the study and summarizes future research directions.

2. Literature Review

Peng et al. (2024) [17] explored the predictive ability of the estimated Glucose Disposal Rate (eGDR) predicting the cardiovascular events and mortality in patient with chronic kidney disease without diabetes. The research design was a prospective cohort study to evaluate clinical and metabolic predictors of cardiovascular events. The findings showed that low eGDR values were highly correlated with increased risk of cardiovascular complication and mortality. The results demonstrated the importance of metabolic risk evaluation in the early cardiovascular risk prediction. The article focused on the need to incorporate metabolic markers into clinical decision-making models.

Kothinti (2024) [18] conducted a detailed analysis of the area of artificial intelligence and its application in predicting diseases and preventative health care. These authors have considered the various ML and DL algorithms used in the preliminary identification of various diseases like cardiovascular and metabolic diseases. The review revealed that AI-based predictive models possess a high degree of diagnostic performance and may be utilized to identify chronic diseases at their early stages. The paper has further noted the significance of data analytics and intelligent decision support system in modern healthcare. The findings revealed the possibility of the transformational character of AI in terms of clinical outcomes and healthcare management enhancement.

Patel et al. (2024) [19] reviewed the best screening protocols in terms of risk prediction of heart failure in adults with diabetes, who have no prior history of atherosclerotic cardiovascular disease. The research involved the use of pooled cohort data in examining clinical markers of heart failure development. The findings showed that the early screening with both metabolic and cardiovascular predictors enhanced future events of cardiac outcomes. The discussion has demonstrated the relevance of preventive screening strategies in diabetic groups. The research added up to more risk stratification and preventive cardiovascular care.

Jose et al. (2024) [20] researched the idea of implementing ML-based methods to health management of cardiovascular in diabetic individuals. These authors research came up with predictive models that can be used to detect cardiovascular risk based on clinical and lifestyle characteristics. The data concerning patients were processed with the help of the ML algorithms and formed the personal intervention plans. This data showed that predictive systems based on smart learning were proved to be extremely helpful in terms of identifying the danger and developing a treatment program at the initial level. The paper has highlighted the ability of the data-driven healthcare systems to be implemented to improve the management of cardiovascular diseases.

Khan et al. (2024) [21] proposed cardiovascular disease detection networks, which are based on ensemble and called EnsCVDD-Net and BlCVDD-Net, which identify heart diseases with high accuracy. The study involved

the combination of a number of ML models on ensemble and combining methodology to optimize the predictive performance. The testing tests showed that the proposed framework was more accurate and solid as compared to conventional classifiers. The model could describe complex associations between clinical risk factors quite well. The findings validated the effectiveness of ensemble learning with regard to the prediction of cardiovascular disease through the task.

Sampath et al. (2024) [22] developed an effective architecture of diabetes prediction on ensemble ML algorithms and Synthetic Minority over-Sampling Technique (SMOTE). The study entailed the problem of class imbalance that is used by medical databases. Ensuring the accuracy of prediction and model stability were accomplished by using a number of ensemble classifiers. The results of the experiment indicated that the mixture of SMOTE and ensemble learning presented significant results in diabetic prediction. The importance of balanced data treatment in healthcare analytics has been brought out by the works of these authors.

The new type of ensemble-based DL-based clinical decision support system proposed by Al Reshan et al. (2024) [23] could help predict diabetes. The model was a combination of different DL models to increase the accuracy of diagnosis and reliability. The proposed system was used to assess the clinical features and offer predictive outcomes in diabetes diagnosis. The prediction using experimental evaluations was found to be superior to the conventional ML. As has been pointed out in the paper, ensemble DL has been effective in the development of intelligent healthcare decision support systems.

Kim et al. (2024) [24] investigated the correlation between Non-Alcoholic Fatty Liver Disease (NAFLD) and cardiovascular disease in patients having type 2 diabetes. The authors used a segment of national population-based data to analyse the results of long-term health. The results showed that diabetic patients with NAFLD were under a significantly high risk of cardiovascular disease and deaths. The findings put into focus the consequences of the metabolism disorders on the cardiovascular health. It was stressed in the paper that it is necessary to monitor and preventively treat high-risk patients at this early phase of development.

Kwiendacz et al. (2025) [25] designed a research based on ML to detect the key adverse cardiac events among diabetic patients with chronic kidney disease. The Silesia Diabetes-Heart Project was also the data used in the study where predictive models were designed using the clinical and biochemical variables. It was done with a series of ML algorithms that were tested in order to determine which prediction strategy is the best. The results indicated that the ML models have a strong contribution to predicting cardiovascular complications. The paper has highlighted the importance of predictive analytics using data of high-risk patients with diabetes.

The research of different heart disease prediction ML algorithm was performed by Rimal et al. (2025) [26]. The study compared the Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbor (KNN)and Random Forest (RF)classifiers through their cross-validation. The results of the experiment proved that ensemble and tree-based models possessed a great degree of successful prediction as compared to traditional models. The article provided information on the value of ML in analyzing clinical cardiovascular data. The paper has assisted in formulating suitable algorithms to be adopted in the prediction of heart diseases in an appropriate manner.

Table 1: Summary of State of the Art ML Approaches for Cardiovascular Disease Prediction in Diabetic Patients

Author & Year	**Objective**	**Method / Model Used**	**Dataset**	**Advantage**	**Disadvantage**
Joel, et al. (2025) [27]	To analyze different imputation techniques for handling missing values in healthcare datasets.	KNN Imputation and comparative imputation methods	Healthcare diagnostic dataset	KNN imputation effectively restored missing medical attributes and improved dataset completeness.	Performance decreases with very large datasets and high dimensional features.
Jarmakovica, A. (2025) [28]	To improve healthcare data quality through ML-based preprocessing strategies.	PCA-based dimensionality reduction and data quality evaluation	Healthcare data repository	PCA effectively reduced redundant features and improved data usability for predictive models.	Important clinical information may be lost if too many dimensions are removed.
Bairy, et	To enhance	Isolation	Healthcare	Isolation Forest	Requires careful

al. (2024) [29]	healthcare data integrity by detecting abnormal records.	Forest (Unsupervised anomaly detection)	dataset	efficiently detects anomalous medical records and improves dataset reliability.	parameter tuning and may misclassify rare but valid medical cases as anomalies.
M. D. Borges (2025) [30]	To identify heart disease risk in diabetic patients using data mining and ML techniques.	LRand RF	Clinical cardiovascular dataset	Ensemble models improved prediction accuracy and effectively handled nonlinear clinical features.	The approach relied on limited clinical attributes and did not incorporate DL or temporal patient data.
N. Nasution et al. (2025) [31]	To compare the performance of multiple ML models for heart disease prediction.	Logistic Regression, RF, SVM, KNN	UCI Heart Disease Dataset	Provided a comprehensive comparison of classical ML algorithms for cardiovascular prediction.	The dataset size was relatively small and may limit generalization to real-world clinical environments.
K. Cao et al. (2025) [32]	To enhance cardiovascular disease prediction using optimized FS techniques.	PSO-XGBoost	Cardiovascular clinical dataset	Feature optimization using Particle Swarm Optimization improved model performance and classification accuracy.	The optimization process increased computational complexity and training time.
I. Rehan & M. U. Rehman (2025) [33]	To develop a hybrid ML framework for cardiovascular risk prediction in diabetic patients.	Hybrid ML Model	Diabetes cardiovascular dataset	Hybrid models combined the strengths of multiple algorithms, improving predictive robustness.	Model interpretability and computational requirements were relatively higher compared to single models.
C. Xu et al. (2025) [34]	To develop and validate a ML model for cardiovascular risk prediction in type-2 diabetic patients.	ML Risk Prediction Model	Type-2 diabetes clinical dataset	The model effectively identified high-risk patients and supported early clinical decision-making.	The framework required large clinical datasets and lacked integration of real-time patient monitoring data.

The summary of Table 1 presents the recent studies which used the use of ML and hybrid models in cardiovascular disease predictors of diabetic patients. The studies mentioned above disclosed that the application of the RF, XGBoost and hybrid learning models contributed to the improvement of the degree of predictive value based on the examination of clinical and metabolic aspects. Not all of these approaches, though, came with disadvantages of increased computational complexity, specific datasets, or less interpretability, which would prefer more efficient predictive frameworks.

3. Proposed Methodologies

This research aims pre processing, and FS to effectively perform the diabetes dataset. First, the A-SPP purges the dataset by eliminating duplicates, treating missing values, identifying outliers. One of the novelties of A-SPP is its clinical-conscious feature tagging, in which attributes undergo analysis using their medical significance and consistency between records of different patients. Following the preprocessing phase, the HFEF does the selection of features by considering feature stability, predictive contribution and redundancy. This algorithm only keeps very informative attributes and eliminates unstable or irrelevant attributes. The originality of this stage is the incorporation of the clinically directed preprocessing and stability-based feature analysis, resulting in a clean and refined dataset of precise diabetes prediction.

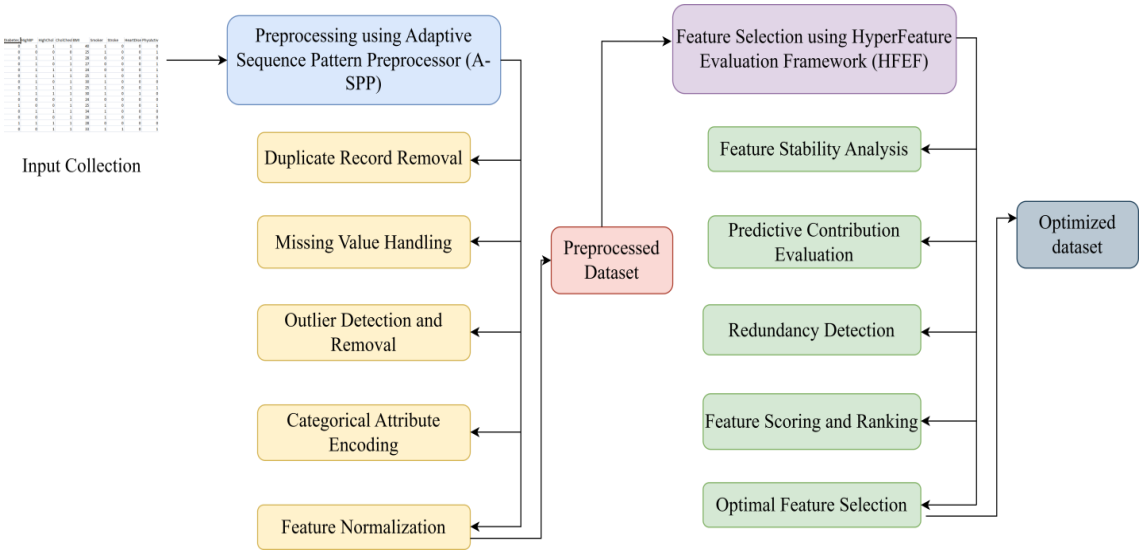


Figure 1: Proposed A-SPP and HFEF Based Heart Disease Prediction Framework

Figure 1 demonstrates the entire process of the proposed framework of heart disease prediction of diabetic patients. It starts with the acquisition of datasets, which is then proceeded by preprocessing A-SPP, FS HFEF and resultant creation of an optimal feature dataset. Last but not the least, the chosen features are predicted and evaluated by performance.

3.1 Dataset Description

Dataset 1: Diabetes Health Indicators Dataset (https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset) is a huge publicly available health dataset based on the Behavioral Risk Factor Surveillance System (BRFSS) survey of 2015 in the U.S. health authorities. It has about 253,680 records of patients with 21 health-related predictive variables and a single target variable of whether diabetic or not. Medical, lifestyle and demographic data contained in the dataset includes BMI, blood pressure, cholesterol, smoking habits, physical activity, history of heart disease, and overall health conditions. The dataset is popular in ML-based research in the field of diabetes and cardiovascular risk prediction due to its large size and a variety of health features.

Dataset 2: The Diabetes Dataset (https://www.kaggle.com/datasets/kandij/diabetes-dataset) is a popular medical dataset of diabetes prediction with the help of ML methods. It has 768 records of patients and 8 clinical attributes and one target variable which is whether a patient has diabetes or not. Other key physiological features that are included in the dataset are the glucose level, blood pressure, insulin level, BMI, age, and the count of pregnancies, among other vital features of diabetes diagnosis. It was first assembled by the National Institute of Diabetes and Digestive and Kidney Diseases and is usually applied to the creation of predictive healthcare models.

3.2 Preprocessing using Adaptive Sequence Pattern Preprocessor (A-SPP)

Adaptive Sequence Pattern Preprocessor (A-SPP) is the program that is aimed at preprocessing the diabetes data to remove inconsistencies and enhance the quality of the data prior to the training of the models. The algorithm first scans the data set to detect and eliminate duplicate records of the patients that can result in bias in the prediction. The approach works on the elimination of duplicates after which missing values are filled which are commonly experienced in medical datasets because of clinical measurements which may not be complete. Such incomplete records are addressed using statistical imputation to maintain the data distribution in general but still ensure clinical reliability.

Thereafter, A-SPP carries out outlier detection with a view of detecting abnormal values that are far apart the expected physiological limits of patient characteristics including glucose level, blood pressure, and body mass index. Eliminating these extreme observations aids with eliminating distorted learning behavior in predictive models. The preprocessing phase then transforms the categorical clinical features into number based representation with the help of label encoding, which allows the ML algorithms to manipulate the data.

In order to achieve a balanced learning in each of the attributes, the algorithm uses feature normalization that linearly scales all the features into a standard numerical range. This stops the attributes with a high magnitude to

dominate the training process. The unique aspect of A-SPP is that the tagging mechanism of clinical-aware features allows the analysis of medical attributes in terms of their relevance and consistency in patterns across patient records. With such adaptive form of tagging, during preprocessing, prioritization of clinically meaningful features is achieved, allowing meaningful medical patterns to be maintained. Consequently, A-SPP produces clean, structured and clinically consistent data, which enhances reliability and predictive accuracy of the next heart disease predictive model.

$$\boldsymbol{D} = \{(\boldsymbol{x_i}, \boldsymbol{y_i})\}_{\boldsymbol{i=1}}^{\boldsymbol{N}} \quad (1)$$

Where, $\boldsymbol{D}$ is whole dataset, $\boldsymbol{x_i}$ is feature vector of the $\boldsymbol{i}$-th patient record,$\boldsymbol{y_i}$is heart disease risk class label, $\boldsymbol{N}$ is total number of samples. Equation (1) is the formatted pattern of data with every patient record being composed of various clinical characteristics and the target label.

$$\boldsymbol{x_i} = (\boldsymbol{f_{i1}}, \boldsymbol{f_{i2}}, \boldsymbol{f_{i3}}, \cdots, \boldsymbol{f_{im}}) \quad (2)$$

Where, $\boldsymbol{f_{ij}}$ is value of the $\boldsymbol{j}$-th feature in the $\boldsymbol{i}$-th record. $\boldsymbol{m}$ is number of features. The Each patient record in equation (2) is a representation of the record as a combination of various medical variables like glucose level, BMI, blood pressure, insulin, and age as a vector.

$$\boldsymbol{sim}(\boldsymbol{x_i}, \boldsymbol{x_j}) = \sum_{\boldsymbol{k=1}}^{\boldsymbol{m}} \boldsymbol{\delta}(\boldsymbol{f_{ik}}, \boldsymbol{f_{jk}}) \quad (3)$$

Where, $\boldsymbol{sim}(\boldsymbol{x_i}, \boldsymbol{x_j})$ is similarity score between records. $\boldsymbol{f_{ik}}, \boldsymbol{f_{jk}}$ are feature values. $\boldsymbol{\delta}$ is matching function.The reason is that in equation (3) the similarity score between two records with the same feature values is the sum of the number of attributes which means that the records are duplicated.

$$\boldsymbol{D_c} = \boldsymbol{D} - \{\boldsymbol{x_i} | \boldsymbol{sim}(\boldsymbol{x_i}, \boldsymbol{x_j}) = \boldsymbol{m}, \boldsymbol{i} \neq \boldsymbol{j}\} \quad (4)$$

Where, $\boldsymbol{D_c}$ is dataset after duplicate removal, $\boldsymbol{m}$ is total features. Equation (4) is deleted when records with the same attributes are deleted in order to remove the redundant patient records.

$$\boldsymbol{f_{ij}} = \begin{cases} \boldsymbol{f_{ij}}, & \boldsymbol{f_{ij}} \neq \emptyset \\ \boldsymbol{\mu_j}, & \boldsymbol{f_{ij}} \neq \emptyset \end{cases} \quad (5)$$

$$\boldsymbol{\mu_j} = \frac{\boldsymbol{1}}{\boldsymbol{N}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{f_{ij}} \quad (6)$$

Where, $\emptyset$ is missing value, $\boldsymbol{\mu_j}$ is mean value of feature $\boldsymbol{j}$. Lacking medical measurements in equation (5) and (6) are substituted with a mean of the corresponding attribute to use in ensuring consistency in the dataset.

$$\boldsymbol{IQR} = \boldsymbol{Q_3} - \boldsymbol{Q_1} \quad (7)$$

Where, $\boldsymbol{Q_1}$ is first quartile, $\boldsymbol{Q_3}$ is quartile. Equation (7) is the used to calculate interquartile range. The value is considered an outlier if

$$\boldsymbol{f_{ij}} < \boldsymbol{Q_1} - \boldsymbol{1.5}\,(\boldsymbol{IQR}) \quad \text{or} \quad \boldsymbol{f_{ij}} > \boldsymbol{Q_3} + \boldsymbol{1.5}\,(\boldsymbol{IQR}) \quad (8)$$

Extreme physiological values that lie outside the normal distribution are identified and eliminated using equation (8) as a measure of preventing bias to the model.

$$\boldsymbol{E}(\boldsymbol{c_k}) = \boldsymbol{z_k} \quad (9)$$

Where, $\boldsymbol{c_k}$ is categorical class, $\boldsymbol{z_k}$ is encoded numerical value. Attributes like gender or smoking status are converted into numerical labels and represented in equation (9) they are compatible with ML.

$$\boldsymbol{x'} = \frac{\boldsymbol{x} - \boldsymbol{x_{min}}}{\boldsymbol{x_{max}} - \boldsymbol{x_{min}}} \quad (10)$$

Where, $\boldsymbol{x}$ is original feature value, $\boldsymbol{x_{min}}$ is minimum feature value, $\boldsymbol{x_{max}}$ is maximum feature value, $\boldsymbol{x'}$ is normalized value. To enhance convergence in the model, Min-max normalization is performed by equation (10) to scale all the attributes between 0 and 1.

$$\boldsymbol{R_j} = \frac{\boldsymbol{1}}{\boldsymbol{N}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{w_i} \boldsymbol{f_{ij}} \quad (11)$$

Where, $\boldsymbol{R_j}$ is clinical relevance score of feature $\boldsymbol{j}$, $\boldsymbol{w_i}$ is importance weight of patient record, $\boldsymbol{f_{ij}}$ is feature value. The medical significance of each attribute is estimated in equation (11), where the weighted attribute contribution in patient records is measured.

$$\boldsymbol{P_j = \frac{1}{N}\sum_{i=1}^{N}(f_{ij} - \mu_j)^2} \qquad (12)$$

Where, $\boldsymbol{P_j}$ is pattern consistency score, $\boldsymbol{\mu_j}$ is mean of feature $\boldsymbol{j}$. Pattern consistency in equation (12) assesses the stability of every clinical feature in the data set.

$$\boldsymbol{C_j = \alpha R_j + \beta P_j} \qquad (13)$$

Where, $\boldsymbol{C_j}$ is final clinical importance score, $\boldsymbol{\alpha and \beta}$ are weighting coefficients. The equation (13) takes medical relevance and statistical consistency to ascertain the significance of every attribute.

$$D_p = \{x_i' | C_j > \theta\} \qquad (14)$$

Where, D_p is preprocessed dataset, C_j is feature importance score, θ is threshold value. Only the clinically significant features are kept in equation (14), the resulting dataset is refined in the subsequent stage of prediction.

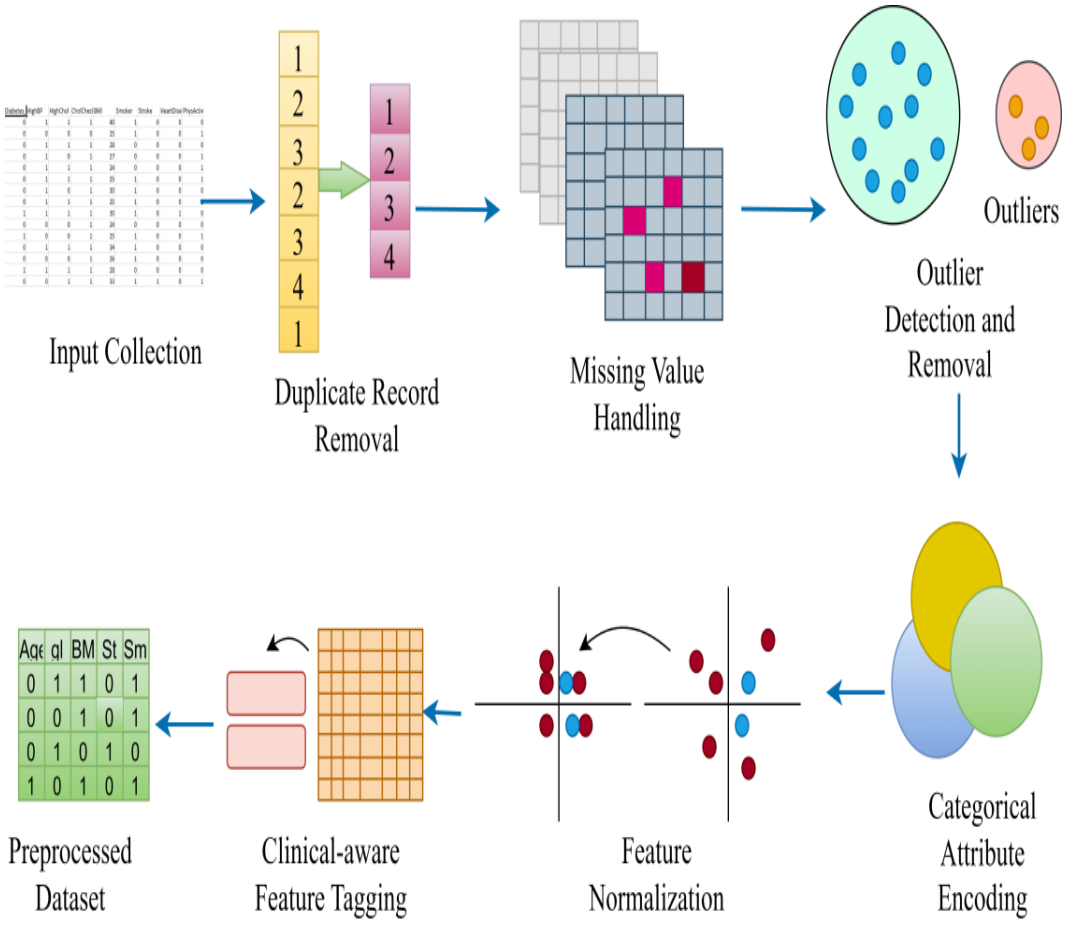


Figure 2: Preprocessing Workflow Architecture

Figure 2 shows the A-SPP to enhance the quality of datasets. It does duplicate removal, missing values, finding outliers, categorical encoding, and feature normalization. This step secures accurate and homogenous data to make competent analysis and forecasts.

Algorithm 1: Adaptive Sequence Pattern Preprocessor (A-SPP)

```
Input raw dataset D
Initialize empty dataset Dclean
Remove duplicate records
For each record xi in D
    If xi is not duplicated in Dclean
        Add xi to Dclean
    End If
End For
Handle missing values
For each feature fj in Dclean
   If value is missing
      Replace with mean or median of feature fj
  End If
End For
Detect outliers
For each feature fj
  Compute Q1, Q3 and IQR
   For each sample xi
     If fj < Q1 − 1.5(IQR) OR fj > Q3 + 1.5(IQR)
        Mark value as outlier
        Replace with feature median
     End If
  End For
End For
Encode categorical attributes
For each categorical feature cj
   Apply label encoding
End For
Normalize features
For each feature fj
   Apply min–max normalization
End For
Perform clinical-aware feature tagging
For each feature fj
   Compute medical relevance score Rj
   Compute pattern consistency score Pj
   Compute clinical score Cj
End For
Construct final dataset D_p
Return D_p
Output: Preprocessed dataset D_p
```

The A-SPP is a tool that cleans up the raw diabetes data that is used by executing a series of preprocessing tasks that include removal of duplicates, imputation of missing data, and detection of outliers so as to enhance the quality of the data. It also encodes categorical attributes into numerical values in the form of label encoding and normalizes features that are on a scale. Lastly, the algorithm uses clinical-aware relevance scores to rank sets of clinical-aware features according to medical importance and consistent patterns and creates a clean and structured dataset that can be used to analyze further.

3.3 Feature Selection using Hyper Feature Evaluation Frameworks

The dataset can still be having irrelevant, unstable, or redundant attributes that can decrease the prediction performance even after preprocessing. In a bid to solve this problem, HFEF is proposed to determine the most pertinent and consistent clinical features in predicting heart disease in diabetic patients. The primary goal of HFEF is to enhance model efficiency through keeping its features that make a substantial contribution to prediction and eliminate noise or repetitive variables. The first step taken by the framework is feature stability analysis that measures the consistency of each attribute across the data set. Attributes whose values vary large or have an irregular occurrence are unstable and will be attributed a lesser stability score. Preferably, stable

features are needed since they add more predictive power to predictive modeling. Then, the predictive contribution analysis is conducted in order to quantify the association between each of the features and the target variable. This step defines the strength of a feature in the prediction of the risk of heart disease. Attributes that have higher predictive values are said to be more informative and they are prioritized when the process of selecting is carried out. The next process is redundancy detection done by HFEF to detect highly correlated attributes that convey similar information. In cases where two or more features have a high correlation, the removal of one of them is done to eliminate the repetition of information and also to minimize the level of computation. These evaluation criteria are then combined into a composite feature score within the framework, which is made up of stability, predictive contribution and redundancy measures. All features are graded based on a weighted score of the relative importance of the feature. Attributes with scores above a predetermined value are chosen as the best subset. This multi-criteria evaluation strategy, used by HFEF, has been placed in a good position to decrease the dimensionality but maintains clinically useful characteristics. The resulting feature subset increases learning ability of the prediction model and the accuracy and strength of predicting heart diseases among diabetic patients.

$$\boldsymbol{F} = \{\boldsymbol{f_1}, \boldsymbol{f_2}, \boldsymbol{f_3}, \dots, \boldsymbol{f_m}\} \qquad (15)$$

Where, $\boldsymbol{F}$ is set of features, $\boldsymbol{m}$ is total number of attributes. Equation (15) defines the set of all candidate features available for evaluation.

$$\boldsymbol{PC_j} = \frac{\sum_{i=1}^{N}(f_{ij}-\mu_j)(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{N}(f_{ij}-\mu_j)^2}\sqrt{\sum_{i=1}^{N}(y_i-\bar{y})^2}} \qquad (16)$$

Where, $\boldsymbol{PC_j}$ is predictive contribution of feature $\boldsymbol{j}$, $\bar{\boldsymbol{y}}$ is mean of target variable. The strength of relationship between a feature and the outcome of a prediction is tested using equation (16).

$$\bar{\boldsymbol{y}} = \frac{\boldsymbol{1}}{\boldsymbol{N}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{y_i} \qquad (17)$$

Equation (17) computes the average class value for correlation calculation.

$$\boldsymbol{R_{jk}} = \boldsymbol{corr}(\boldsymbol{f_j}, \boldsymbol{f_k}) \qquad (18)$$

Where, $\boldsymbol{R_{jk}}$ is redundancy between feature $\boldsymbol{j}$ and $\boldsymbol{k}$. Features with a high correlation (equation (18)) are redundant and similar to each other.

f_k is removed if $|\boldsymbol{R_{jk}}| > \gamma$ (19)

Where, γ is redundancy threshold. In equation (19) if correlation exceeds the threshold, one feature is eliminated.

$$FI_j = \alpha S_j + \beta \boldsymbol{PC_j} + \lambda \mathrm{R_j} \qquad (20)$$

Where, FI_j is final feature importance score, S_j is stability score, $\boldsymbol{PC_j}$ is predictive contribution, $\mathrm{R_j}$ is redundancy factor, α, β, λ are weighting parameters. Equation (20) integrates multiple evaluation criteria into a single importance score.

$$Rank(f_j) = Sort\ (FI_j) \qquad (21)$$

In equation (21) the features are ranked according to their importance scores.

$$F_o = \{f_j | FI_j > \theta\} \qquad (22)$$

Where, F_o is optimal feature subset, θ is selection threshold. In Equation (22) only features with high importance scores are retained.

$$D_o = \{(x_i^{F_o}, y_i)\}_{i=1}^{N} \qquad (23)$$

Where, D_o is dataset containing selected features. This dataset (equation (23) is used for the final prediction model.

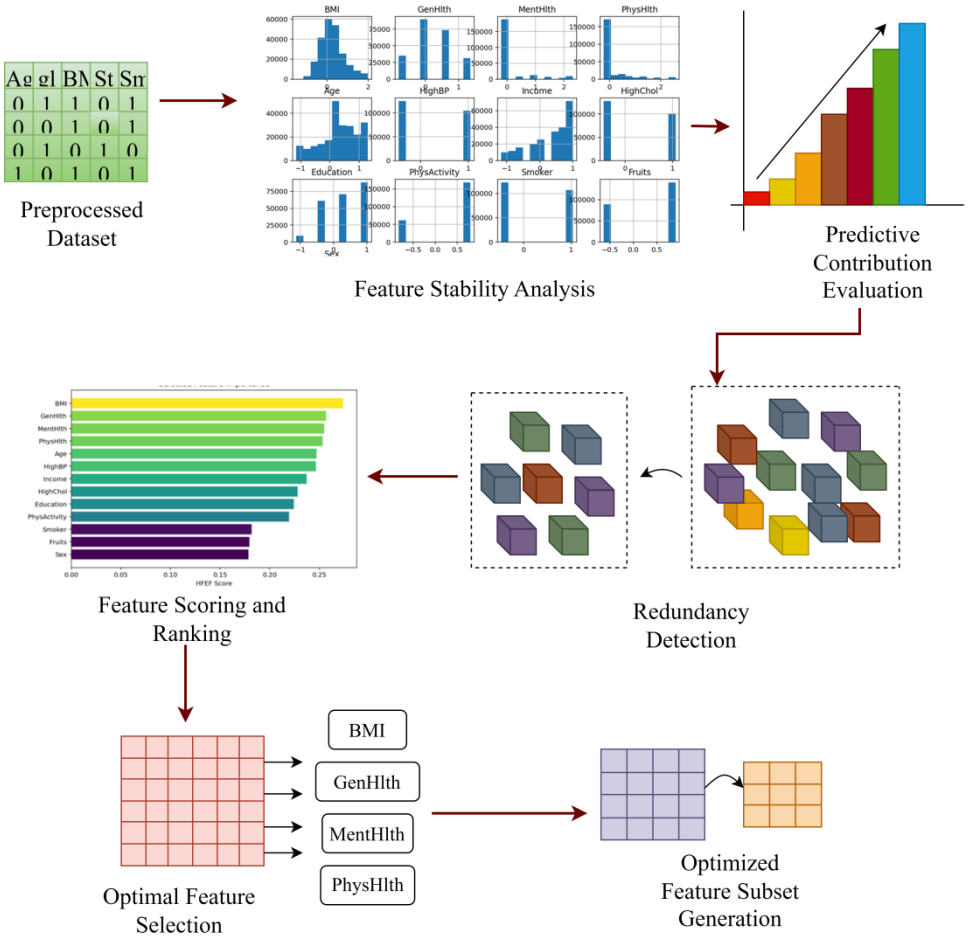


Figure 3: Feature Selection Workflow Architecture

Figure 3 describes the HFEF to be used in the identification of the most informative attributes. It approximates the stable, predictive, and redundant features and ranks the features and picks the most pertinent clinical features. The dimension of datasets is pruned in this process in addition to improving predicting proficiency.

Algorithm 2: HyperFeature Evaluation Framework (HFEF)

Input preprocessed dataset Dp
Extract feature set F = {f1, f2, … , fm}
Compute feature stability
For each feature fj in F
  Compute mean μj
  Compute stability score Sj
End For
Evaluate predictive contribution
For each feature fj
  Compute correlation PCj between fj and target variable
End For
Detect feature redundancy
For each pair of features (fj , fk)
  Compute correlation Rjk
  If |Rjk| > threshold γ
  Mark one feature as redundant
  End If
End For
Compute final feature importance
For each feature fj
  Calculate FIj = αSj + βPCj − λRj

End For
Rank features based on FIj
Select features with FIj greater than threshold θ
Construct optimal feature subset F_o
Return F_o
Output: Optimal feature subset F_o

The HFEF selects the most ranked features of the processed data in order to improve accuracy in prediction and reduction in computational costs. It also compares every feature according to the stability analysis, predictive contribution in comparison to the target feature and redundancy detection between correlated features. The algorithm ranks the attributes in order of cumulative features importance score and then selects the optimal subset of attributes that adds significant predictive value of heart disease in diabetic patients.

3.4 Computational Complexity Analysis

The computational complexity of the proposed A-SPP and HFEF algorithms is analyzed to ascertain the efficiency and scalability of the algorithms. Where N is the size of the overall dataset of the patient records and MMM represents the number of clinical features. The cost of processing each processing stage is calculated by the quantity of operations required to process the set of data.

During A-SPP preprocessing phase, duplicate detection may be done by performing comparisons between records in the dataset, with a worst case of O(N 2) operations. The missing value imputation and normalization algorithms are O(NM) to compute on all features of all samples. Similarly, interquartile range technique of outliers detection is a process of scanning all the feature values and its complexity is O(NM). Going through all feature values once is also required to the categorical encoding process and introduces another O(NM). Therefore, the most time-consuming part of the preprocessing phase is the duplicate detection and can be estimated as indicated in equation (23):

$$T(ASPP) = O(N^2) + O(NM) \qquad (23)$$

Where T (n A B C 4) d is the overall computational cost of the preprocessing algorithm. In order to achieve the step of FS in the case of the HFEF, feature stability analysis would be the evaluation of the variation of the entire sample of a feature which can be a O(NM) complexity. To compute the correlation between each feature and target variable, predictive contribution analysis is computed to take O(NM) operations, as well. The redundancy detection algorithm performs a pairwise correlation test on the features providing a complexity of O (M 2). The sorting processes of the feature ranking and threshold selection require O(M log M) sorting. Thus, the overall complexity of calculations of the FS framework can be expressed into equation (24) below:

$$T(HFEF) = O(NM) + O(M^2) + O(Mlog\,M) \qquad (24)$$

Where $T(HFEF)$represents the cost of computation of the feature evaluation framework. Taking the two stages together the complexity of the entire computation involved in the proposed framework can be stated as represented in equation (25):

$$T(Total) = O(N^2) + O(NM) + O(M^2) + O(Mlog\,M) \qquad (25)$$

It is recommended in the analysis that the specified methodology remains computationally effective with relatively small data and with the huge sized data the dimensionality reduction produced by HFEF decreases the scale of the data significantly consequently, minimizing the cost of computation of the subsequent provided prediction model.

3.5 Parameter Analysis

The parameter analysis is carried out to obtain the notion of the significance of parameters implemented in the A-SPP and HFEF. The right selection of these parameters is the key of successful preprocessing, correct evaluation of features, and enhanced prediction. The proposed framework uses several parameters, which control the stability of the features in terms of prediction of value, elimination of redundancy and thresholding of features selection.

The parameters are the size of the dataset N, the number of features M and the statistical thresholds that are used to identify an outlier among others that influence the performance of the preprocessing stage A-SPP. Sensitivity

Outlier detection is determined by the threshold coefficient of interquartile range (usually 1.5). A higher threshold may leave extreme values in the data set and reduce the threshold may remove potentially useful data samples. Equally, normalization parameters X_{min} and X_{max} are used to estimate the range of scaling used to transform the feature values to a uniform domain.

The final feature importance score calculated in the HFEF FS step is calculated by using weighting coefficients, α, β and λ, in which lambda is a feature stability contribution, feature predictive correlation and a feature redundancy penalty. This redundancy threshold γ is used to determine the scenario where two features are considered to be highly correlated such that the framework can remove redundant attributes. Moreover, the FS threshold 0 is the most important in determining the size of the best feature subset.

The right trade-off between these parameters will ensure that the framework is suitable with regard to the identification of informative clinical features and will also eliminate redundant and unstable attributes. The sensitive parameter selection enhances predictability of the selected feature set and accuracy of the subsequent model of prediction of heart diseases.

Table 2: Key Parameters Used in the Proposed Framework

Parameter	Description
(N)	Total number of patient records in the dataset
(M)	Number of clinical features
(X_{min})	Minimum value of a feature used in normalization
(X_{max})	Maximum value of a feature used in normalization
(α)	Weight assigned to feature stability score
(β)	Weight assigned to predictive contribution score
(λ)	Redundancy penalty coefficient
(γ)	Correlation threshold used to detect redundant features
(θ)	Feature importance threshold for selecting optimal attributes

It is proposed to set up the framework in table 2 to be able to accommodate other databases and ensure that the features that will be selected will provide reliable and clinically significant information that will be used to predict heart disease in diabetic patients.

3.6 Clinical Relevance Justification

The proposed framework uses the clinically significant attributes to ensure that the prediction model can hint the risk of heart diseases amongst diabetic patients in a medical manner. Among the main physiological factors, glucose level, blood pressure, BMI, cholesterol, and age have been pointed out as one of the most important risk factors that lead to the poor cardiovascular health.

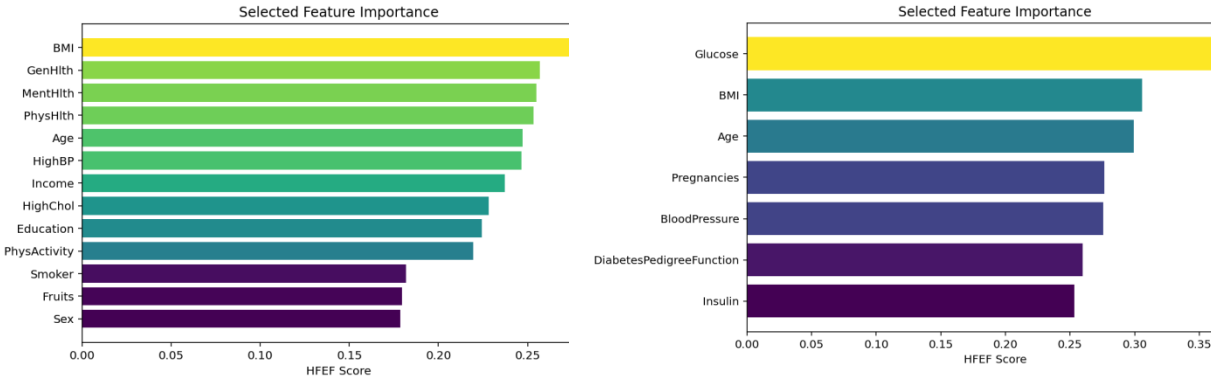
It is in A-SPP preprocessing phase that the features are assigned relevance scores using their medical importance and pattern consistency as compared to the HFEF FS framework which prioritizes these attributes during the selection process. This is an interpretability model attribute that can be trained on clinical guidelines to advance the credibility of the forecasts of heart diseases among diabetic individuals.

4. Results and Discussions

This part reflects the experimental implementation of the suggested scheme of the heart disease prediction of diabetic patients. The A-SPP and HFEF that were proposed were implemented in the Python programming environment. The success of the suggested method is estimated through the assistance of the conventional measures such as accuracy, precision, recall, F1-score, and error indicators. Also, the results are compared with some of the currently existing ML models and demonstrated to be effective and reliable. The obtained results highlight the ability of the proposed method to make the forecast more efficient and provide more accurate risk evaluation.

4.1 Performance Evaluation

The framework effectiveness is then tested in the comparison between the preprocessing and effectiveness of FS. Error metrics such as MSE, RMSE and MAE are used to measure the performance of preprocessing, and accuracy, precision, recall and F1-score are used to measure the efficiency of the features selected in predicting heart diseases more efficiently.

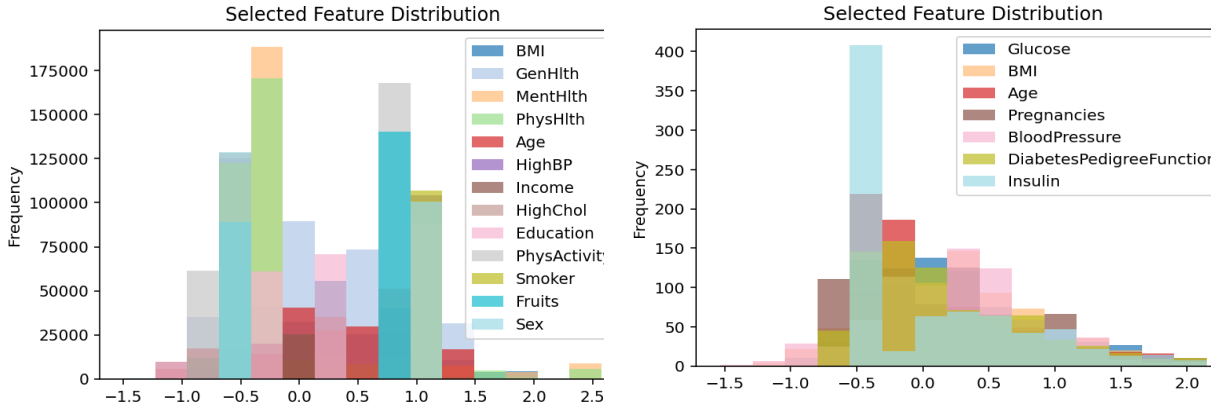


a) Selected Features Importance in Dataset 1

b) Selected Features Importance in Dataset 1

Figure 4: Feature Ranking in Datasets 1 and 2

The scores of the feature importance acquired on the HFEF represented in Figure 4 (a) were acquired by the process of features choice using dataset 1. It shows that, BMI, GenHlth, MentHlth and PhysHlth have the largest effect on the prediction of heart diseases, and this indicates that the clinical variables under question have the most influence in showing the cardiovascular risk in diabetic patients. As Figure 4 (b) indicates, the strongest impact is made by Glucose, BMI, Age, Pregnancies, which implies that clinical variables have the most significant influence on the process of identifying heart diseases in diabetic patients.

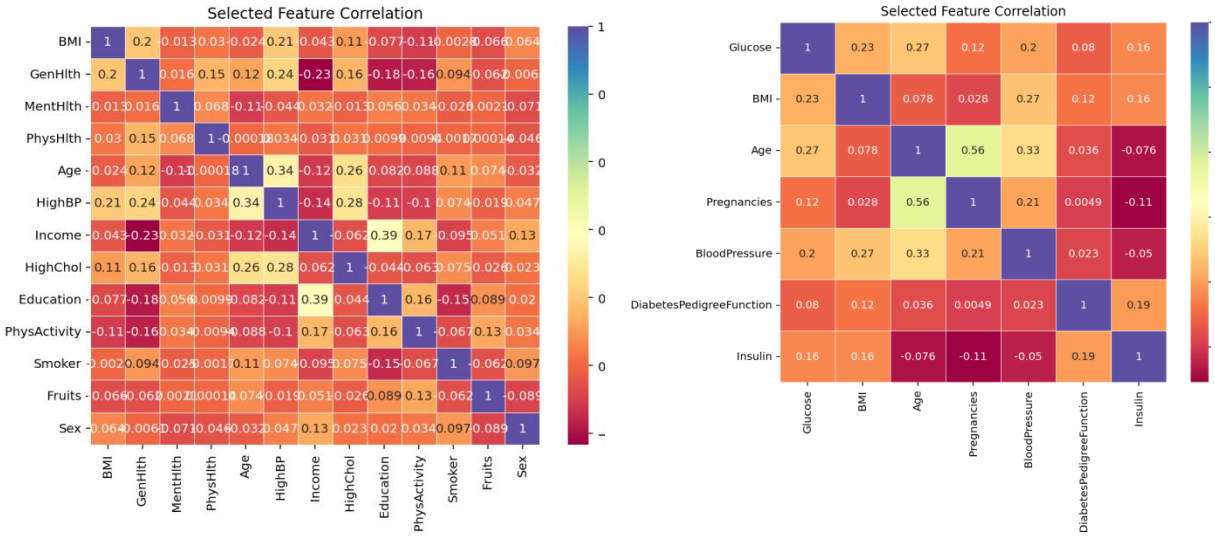


a) Selected Feature Distribution of Dataset 1

b) Selected Feature Distribution of Dataset 2

Figure 5: Feature Distribution of dataset 1 and 2

Figure 5 demonstrates the distribution of the chosen clinical features after postprocessing and selecting features with the proposed framework with the help of dataset 1(a) and dataset 2(b). It presents the distribution of the attributes within the dataset, which demonstrates the balanced and normalized values of features which can be trusted to forecast heart diseases.



a) Dataset 1 Selected Feature Correlation **b) Dataset 2 Selected Feature Correlation**

Figure 6: Feature Correlation of Dataset 1 and 2

The correlation between the selected features and represents the relationships between clinical attributes of dataset 1 and dataset 2, respectively, presented in figure 6 (a) and (b), respectively. It demonstrates that most of the features have low to medium correlations that are indicative of the effectiveness of the HFEF approach in removing redundancy and preserving the information utilized in predicting heart diseases.

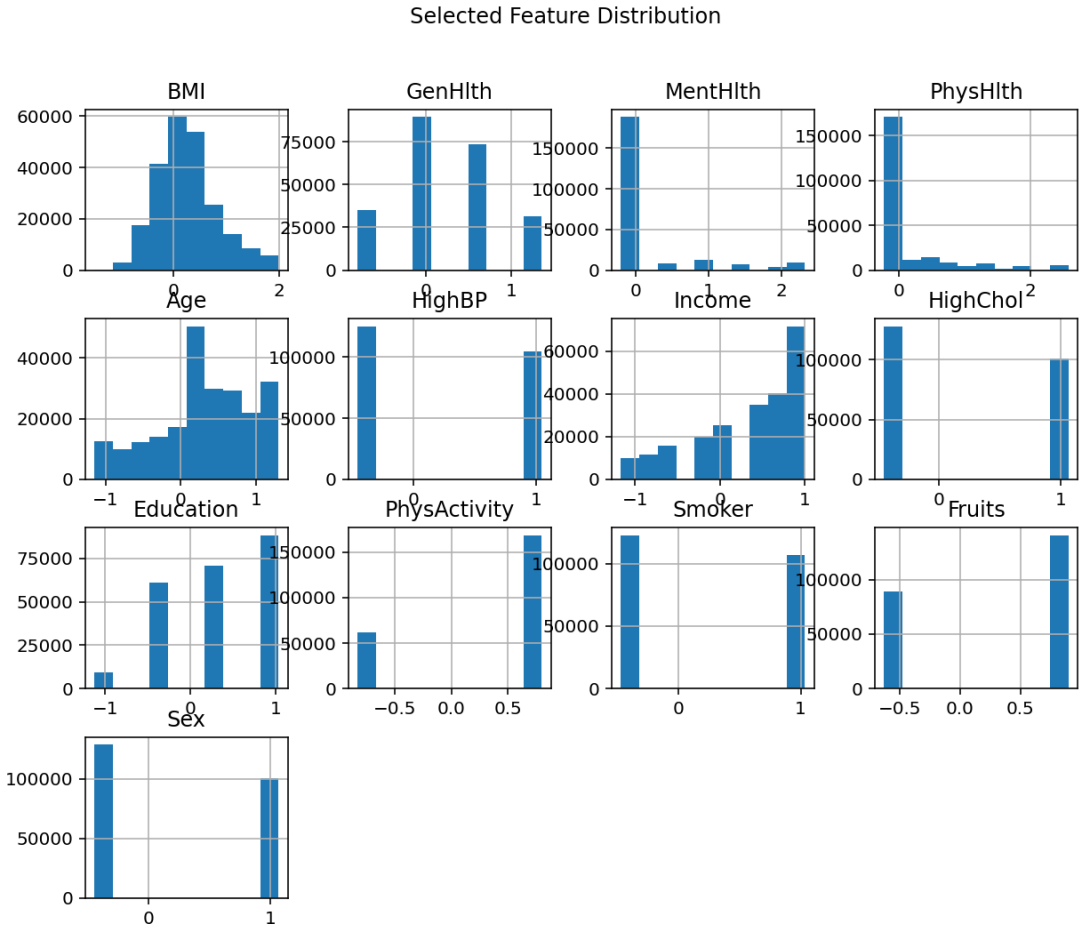


Figure 7: Feature Distribution Histogram using Dataset 1

The distribution of the chosen health-related attributes in the dataset is shown in Figure 7. Most of the variables such as HighBP, HighChol, Smoker, PhysActivity, Fruits and Sex are binary (0 or 1) and have two separate bars meaning the presence or the absence of the condition or behavior. The spreads of continuous or ordinal variables like BMI, Age, Income, Education, GenHlth, MentHlth, and PhysHlth demonstrate a difference in health status, demographic variables and lifestyle of the individuals present in the dataset.

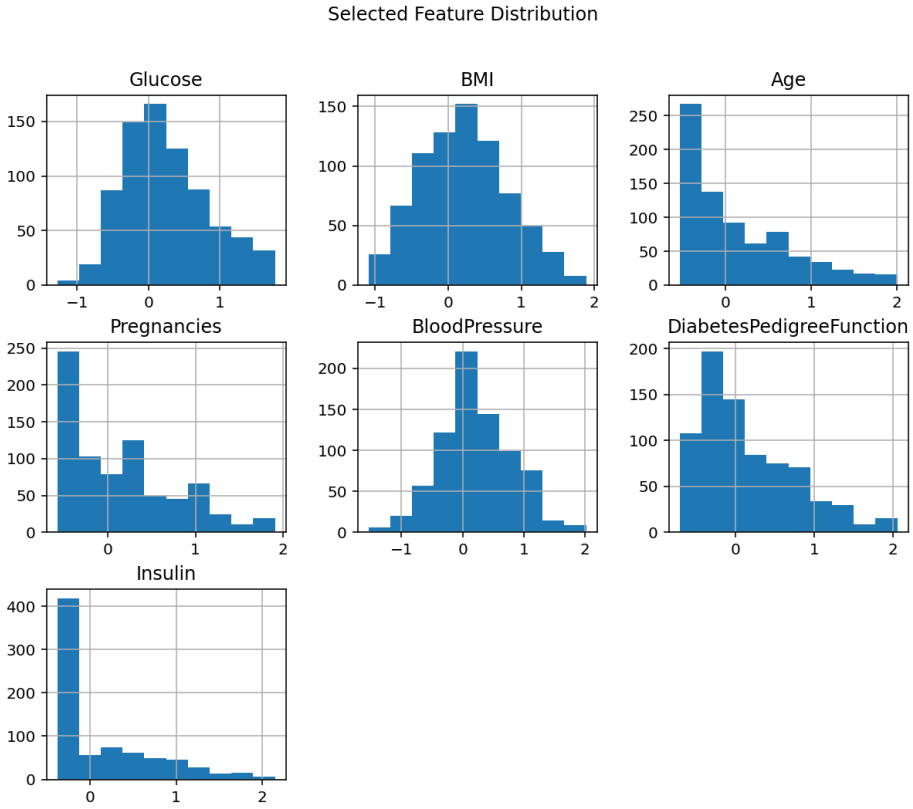
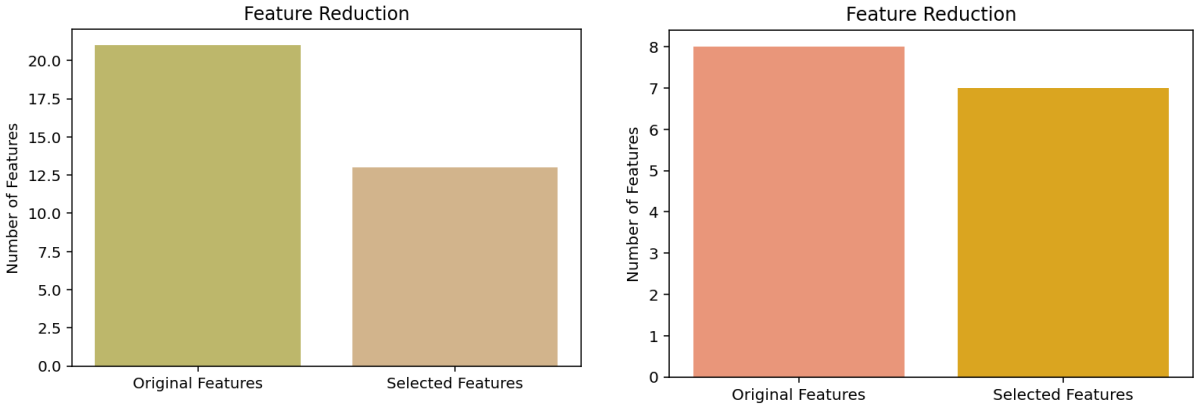


Figure 8: Feature Distribution Histogram using Dataset 2

Figure 8 show distribution of the major medical features including Glucose, BMI, Age, Pregnancies, BloodPressure, Diabetes, Pedigree Function and Insulin using the dataset. The variables such as Glucose, BMI, and BloodPressure are approximately normal, whereas all the other variables such as Pregnancies, Age, and Insulin are skewed to the right, meaning most of the people will show lower values with lesser numbers occurring in high values.



a) Feature Reduction in Dataset 1

b) Feature Reduction in Dataset 2

Figure 9: Feature reduction Comparison of Dataset 1and 2

The reduction of features in the dataset 1 was illustrated in figure 9 (a). Originally, there were 21 features in the dataset, but the dataset was cut down to 13 important features using FS methods. The figure 9 (b) depicts the feature reduction procedure of the dataset 2. The dataset was originally represented by 8 features; however, 7 important features remained by using the methods of FS.

Table 3: Preprocessing methods Performance Comparison using Dataset 1 and Dataset 2

Dataset	**Dataset 1**			**Dataset 2**		
Methods/ Metrics	**MSE**	**RMSE**	**MAE**	**MSE**	**RMSE**	**MAE**
Z-Score Normalization	0.4126	0.6423	0.6015	0.3658	0.6048	0.5662
SMOTE	0.3785	0.6152	0.5741	0.3324	0.5766	0.5389
KNN Imputation	0.3412	0.5841	0.5487	0.2967	0.5447	0.5123
PCA	0.3098	0.5566	0.5215	0.2642	0.5140	0.4827
Isolation Forest	0.2746	0.5240	0.4983	0.2215	0.4706	0.4451
A-SPP (Proposed)	0.2353	0.4850	0.4704	0.1871	0.4326	0.4024

Table 3 determines performance of preprocessing using Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) on two datasets. The smaller the values, the higher are the quality of the data and its preprocessing efficiency. The proposed A-SPP shows the minimal error values in both datasets, proving that it is able to clean, normalize, and optimize the dataset better than the conventional preprocessing methods.

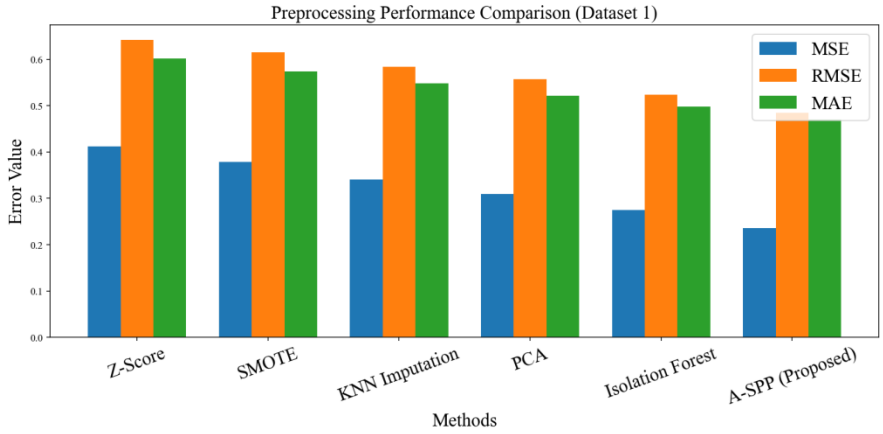


Figure 10: Preprocessing Performance Comparison Chart using Dataset 1

Figure 10 illustrates the preprocessing methods comparison chart using Dataset 1 according to MSE, RMSE, and MAE metrics. In this chart the horizontal line shows the preprocessing methods and the vertical line shows the error values.

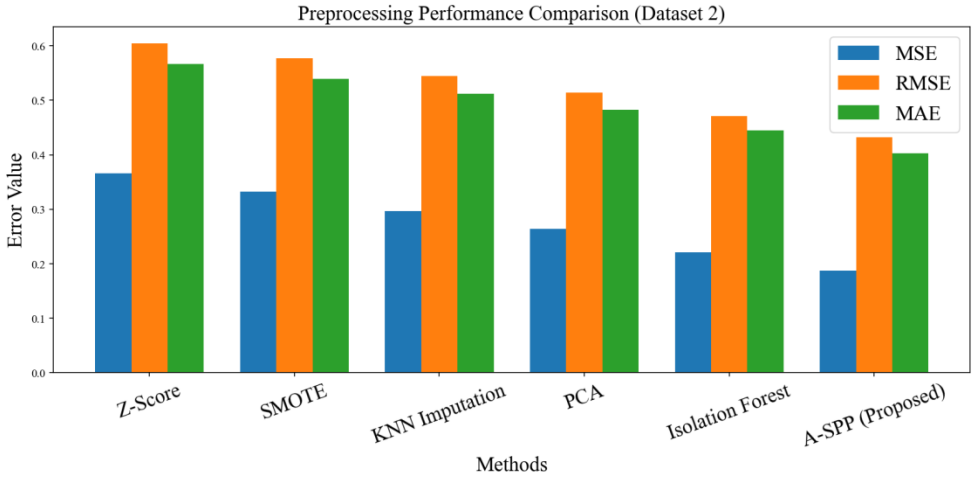


Figure 11: Preprocessing Performance Comparison Chart using Dataset 2

Figure 11 illustrates the preprocessing methods comparison chart using Dataset 2 according to MSE, RMSE, and MAE metrics. In this chart the horizontal lineshows the preprocessing methods and the vertical lineshows the error values.

Table 4: FS Algorithms Performance Comparison

Dataset 1				
	Accuracy	**Precision**	**Recall**	**F1-score**
LR-RF	0.884	0.872	0.865	0.868
SVM	0.901	0.893	0.887	0.890
PSO-XGBoost	0.915	0.907	0.903	0.905
Hybrid ML	0.923	0.916	0.911	0.913
Hybrid ML with LSTM	0.931	0.925	0.919	0.922
HFEF (Proposed)	0.939	0.933	0.928	0.930
Dataset 2				
LR-RF	0.892	0.883	0.878	0.880
SVM	0.908	0.900	0.895	0.897
PSO-XGBoost	0.921	0.915	0.909	0.912
Hybrid ML	0.929	0.922	0.917	0.919
Hybrid ML with LSTM	0.937	0.931	0.926	0.928
HFEF (Proposed)	0.943	0.938	0.932	0.935

Table 4 illustrates the performance comparisons among several FS models, the proposed HFEF-based framework is the most accurate, the most precise, the most recall, and the most F1-score at both dataset 1 and dataset 2. The above-mentioned improvement illustrates that the subset of optimized features generated by HFEF improves the predictive strength of the prediction model and increases the accuracy of heart disease risk prediction in diabetic patients.

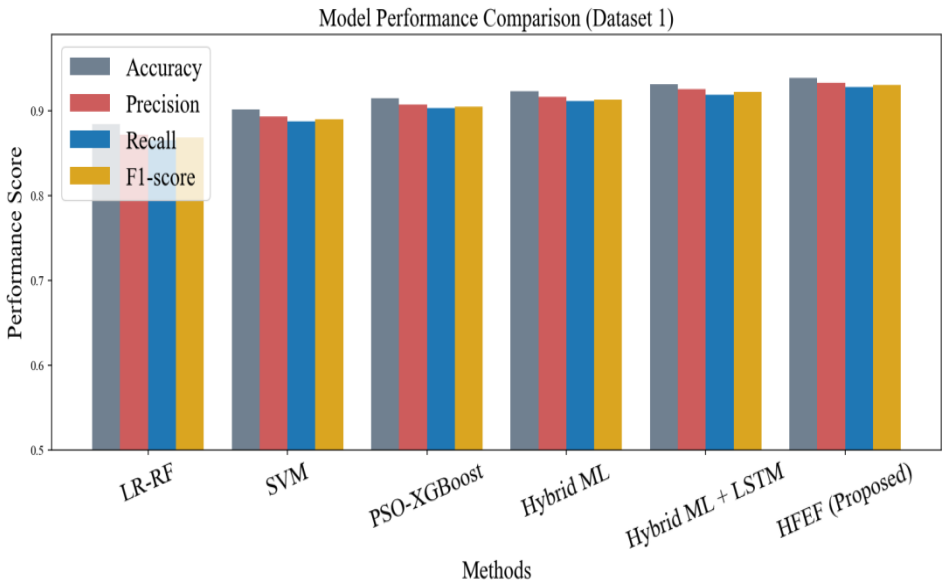


Figure 12:FS Methods Comparison Chart using Dataset 1

Figure 12 shows the performance comparison of FS methods used in heart disease prediction in diabetic patients dataset 1. In this chart the horizontal lineshows the FS methods and the vertical lineshows the performance score.

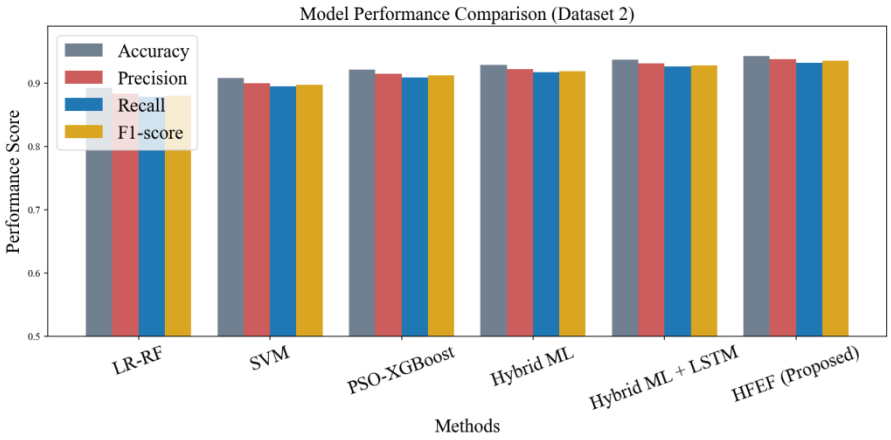


Figure 13: FS Methods Comparison Chart using Dataset 2

Figure 13 shows the performance comparison of FS methods used in heart disease prediction in diabetic patients dataset 2. In this chart the horizontal lineshows the FS methods and the vertical lineshows the performance score.

4.2 Ablation Study

A ablation studyis used to test the individual contribution of each part of the proposed framework, that is, the A-SPP and the HFEF. The analysis presented allows concluding on the influence of each of the modules on the global prediction of heart disease risk in diabetic patients. There are four experimental settings that are tested and these are: a baseline without preprocessing and FS, model with A-SPP preprocessing only, model with HFEF FS only and the proposed framework is a combination of A-SPP and HFEF.

Table 5: Ablation Study Comparison Table

	Accuracy	
Model Configuration	**Dataset 1**	**Dataset 2**
Baseline Model (No preprocessing, No FS)	0.822	0.831
A-SPP Only	0.901	0.906
HFEF Only	0.918	0.920
A-SPP + HFEF (Proposed)	0.939	0.943

Table 5 shows that the A-SPP preprocessing with HFEF FS can improve the prediction performance. The results of the two modules are best when the two are combined, and the model accuracy is 0.939 and 0.943 on Dataset 1 and Dataset 2, respectively. This proves that the integrated framework is effective in enhancing the quality of data and relevance of features that result in more accurate prediction of heart diseases among diabetic patients.

4.3 Discussions

The experimental findings show that the proposed HFEF based framework is superior in its performance in relation to the current models in both datasets. The enhancement is largely seen to be due to the A-SPP that is ideal in cleaning the data by eliminating duplicate data, dealing with missing data, outlier detection and normalization of the feature values. This is a pre-processing step that improves the quality of data and guarantees quality input into the prediction model.

Also, the HFEF enhances the performance of any model by choosing the most informative and stable clinical features and removes overlapping features. Consequently, the suggested method takes maximum accuracy values of 0.939 and 0.943 on Dataset 1 and Dataset 2 correspondingly, and better values of precision, recall, and F1-score. These findings suggest that the given framework is a credible and efficient tool that can be used to predict heart diseases among diabetic patients.

5. Conclusion

The researchproposesnovel model of predicting the risk of heart disease in diabetic patients based on a combined preprocessing and FS strategy. Proposed method The proposed method uses the A-SPP to clean and normalize the dataset by eliminating duplicates, missing values, outliers, and encoding of categorical attributes.

The preprocessing step enhances quality data and provides assured information into the training of the model. After preprocessing, HFEF is also presented to determine the most informative and stable features depending on the redundancy and predictive contribution analysis. The framework improves the prediction model learning ability by choosing the clinically relevant attributes. Experiments with two diabetes-based datasets prove that the suggested model outperforms other known ML models. The framework produced accuracy of 0.939 and 0.943 with better precision, recall and F1-score. The ablation experiment also confirms the fact that the combination of A-SPP and HFEF is important in enhancing prediction performance. In general, the suggested approach offers a valid and effective way of identifying the risk of heart disease in diabetic patients at an earlier stage. The next direction of work will be to use the DL models and the increased scale of clinical data to improve the accuracy of prediction and its practical use even more.

References

1. Rajbhandari, J., Fernandez, C. J., Agarwal, M., Yeap, B. X. Y., &Pappachan, J. M. (2021). Diabetic heart disease: A clinical update. *World journal of diabetes*, *12*(4), 383.

2. Pandey, A., Vaduganathan, M., Patel, K. V., Ayers, C., Ballantyne, C. M., Kosiborod, M. N., ... & Everett, B. M. (2021). Biomarker-based risk prediction of incident heart failure in pre-diabetes and diabetes. *Heart Failure*, *9*(3), 215-223.

3. Zhang, Z., Zhao, L., Lu, Y., Xiao, Y., & Zhou, X. (2024). Insulin resistance assessed by estimated glucose disposal rate and risk of incident cardiovascular diseases among individuals without diabetes: findings from a nationwide, population based, prospective cohort study. *Cardiovascular diabetology*, *23*(1), 194.

4. Fan, R., Zhang, N., Yang, L., Ke, J., Zhao, D., & Cui, Q. (2020). AI-based prediction for the risk of coronary heart disease among patients with type 2 diabetes mellitus. *Scientific reports*, *10*(1), 14457.

5. Butt, U. M., Letchmunan, S., Ali, M., Hassan, F. H., Baqir, A., & Sherazi, H. H. R. (2021). Machine learning based diabetes classification and prediction for healthcare applications. *Journal of healthcare engineering*, *2021*(1), 9930985.

6. Krishnamoorthi, R., Joshi, S., Almarzouki, H. Z., Shukla, P. K., Rizwan, A., Kalpana, C., & Tiwari, B. (2022). [Retracted] A novel diabetes healthcare disease prediction framework using machine learning techniques. *Journal of healthcare engineering*, *2022*(1), 1684017.

7. Febrian, M. E., Ferdinan, F. X., Sendani, G. P., Suryanigrum, K. M., &Yunanda, R. (2023). Diabetes prediction using supervised machine learning. *Procedia Computer Science*, *216*, 21-30.

8. Ogunpola, A., Saeed, F., Basurra, S., Albarrak, A. M., & Qasem, S. N. (2024). Machine learning-based predictive models for detection of cardiovascular diseases. *Diagnostics*, *14*(2), 144.

9. Almazroi, A. A., Aldhahri, E. A., Bashir, S., & Ashfaq, S. (2023). A clinical decision support system for heart disease prediction using deep learning. *IEEE Access*, *11*, 61646-61659.

10. Kee, O. T., Harun, H., Mustafa, N., Abdul Murad, N. A., Chin, S. F., Jaafar, R., & Abdullah, N. (2023). Cardiovascular complications in a diabetes prediction model using machine learning: a systematic review. *Cardiovascular Diabetology*, *22*(1), 13.

11. Oikonomou, E. K., & Khera, R. (2023). Machine learning in precision diabetes care and cardiovascular risk prediction. *Cardiovascular Diabetology*, *22*(1), 259.

12. Ahmad, A., Lim, L. L., Morieri, M. L., Tam, C. H. T., Cheng, F., Chikowore, T., ... &Mathioudakis, N. (2024). Precision prognostics for cardiovascular disease in type 2 diabetes: a systematic review and meta-analysis. *Communications medicine*, *4*(1), 11.

13. Shynu, P. G., Menon, V. G., Kumar, R. L., Kadry, S., & Nam, Y. (2021). Blockchain-based secure healthcare application for diabetic-cardio disease prediction in fog computing. *IEEE Access*, *9*, 45706-45720.

14. Ritchie, R. H., & Abel, E. D. (2020). Basic mechanisms of diabetic heart disease. *Circulation research*, *126*(11), 1501-1525.

15. Kong, D., Chen, R., Chen, Y., Zhao, L., Huang, R., Luo, L., ... & Ding, Y. (2024). Bayesian network analysis of factors influencing type 2 diabetes, coronary heart disease, and their comorbidities. *BMC Public Health*, *24*(1), 1267.

16. Peng, J., Zhang, Y., Zhu, Y., Chen, W., Chen, L., Ma, F., ... & Huang, Z. (2024). Estimated glucose disposal rate for predicting cardiovascular events and mortality in patients with non-diabetic chronic kidney disease: a prospective cohort study. *BMC medicine*, *22*(1), 411.

17. Kothinti, R. R. (2024). Artificial Intelligence in Disease Prediction: Transforming Early Diagnosis and Preventive Healthcare. *Int. J. Acad. Res. Dev*, *9*, 1330-1340.

18. Patel, K. V., Segar, M. W., Klonoff, D. C., Khan, M. S., Usman, M. S., Lam, C. S., ... & Pandey, A. (2024). Optimal screening for predicting and preventing the risk of heart failure among adults with diabetes without atherosclerotic cardiovascular disease: a pooled cohort analysis. *Circulation*, *149*(4), 293-304.

19. Jose, R., Syed, F., Thomas, A., & Toma, M. (2024). Cardiovascular health management in diabetic patients with machine-learning-driven predictions and interventions. *Applied Sciences*, *14*(5), 2132.

20. Khan, H., Javaid, N., Bashir, T., Akbar, M., Alrajeh, N., & Aslam, S. (2024). Heart disease prediction using novel ensemble and blending based cardiovascular disease detection networks: EnsCVDD-Net and BlCVDD-Net. *IEEE Access*, *12*, 109230-109254.

21. Sampath, P., Elangovan, G., Ravichandran, K., Shanmuganathan, V., Pasupathi, S., Chakrabarti, T., ... &Margala, M. (2024). Robust diabetic prediction using ensemble machine learning models with synthetic minority over-sampling technique. *Scientific reports*, *14*(1), 28984.

22. Al Reshan, M. S., Amin, S., Zeb, M. A., Sulaiman, A., Alshahrani, H., Shaikh, A., &Elmagzoub, M. A. (2024). An innovative ensemble deep learning clinical decision support system for diabetes prediction. *IEEE Access*, *12*, 106193-106210.

23. Kim, K. S., Hong, S., Han, K., & Park, C. Y. (2024). Association of non-alcoholic fatty liver disease with cardiovascular disease and all cause death in patients with type 2 diabetes mellitus: nationwide population based study. *Bmj*, *384*.

24. Kwiendacz, H., Huang, B., Chen, Y., Janota, O., Irlik, K., Liu, Y., ... &Nabrdalik, K. (2025). Predicting major adverse cardiac events in diabetes and chronic kidney disease: a machine learning study from the Silesia Diabetes-Heart Project. *Cardiovascular diabetology*, *24*(1), 76.

25. Rimal, Y., Sharma, N., Paudel, S., Alsadoon, A., Koirala, M. P., & Gill, S. (2025). Comparative analysis of heart disease prediction using logistic regression, SVM, KNN, and random forest with cross-validation for improved accuracy. *Scientific Reports*, *15*(1), 13444.

26. Joel, L. O., Doorsamy, W., & Paul, B. S. (2025). A comparative study of imputation techniques for missing values in healthcare diagnostic datasets. *International Journal of Data Science and Analytics, 20*, 6357–6373.

27. Jarmakovica, A. (2025). Machine learning-based strategies for improving healthcare data quality: An evaluation of accuracy, completeness, and reusability. *Frontiers in Artificial Intelligence, 8*, 1621514.

28. Bairy, M., Muniyal, B., & Shetty, N. P. (2024). Enhancing healthcare data integrity: Fraud detection using unsupervised learning techniques. *International Journal of Computers and Applications, 46*(11), 1006–1019.

29. Borges, M. D. (2025). Advancing Deep Learning Insights for Identifying Heart Disease in Diabetic Patients: A Data Mining Approach Using Logistic Regression and Random Forests. *Available at SSRN 5091734*.

30. T M Sathish Kumar, & Shaik Sadulla. (2025). Fractional-Order Mathematical Models for Vibration Analysis in Smart Structural Systems. *Journal of Applied Mathematical Models in Engineering*, *1*(4), 1–8.

31. Nasution, N., Hasan, M. A., & Nasution, F. B. (2025). Predicting Heart Disease Using Machine Learning: An Evaluation of Logistic Regression, Random Forest, SVM, and KNN Models on the UCI Heart Disease Dataset. *IT Journal Research and Development*, *9*(2), 140-150.

32. Cao, K., Liu, C., Yang, S., Zhang, Y., Li, L., Jung, H., & Zhang, S. (2025). Prediction of cardiovascular disease based on multiple FS and improved PSO-XGBoost model. *Scientific Reports*, *15*(1), 12406.

33. Rehan, I., & Rehman, M. U. (2025). Cardiovascular risk prediction in diabetes: a hybrid machine learning approach. *Biomedical Physics & Engineering Express*, *11*(6), 065016.

34. Xu, C., Shi, F., Ding, W., Fang, C., & Fang, C. (2025). Development and validation of a machine learning model for cardiovascular disease risk prediction in type 2 diabetes patients. *Scientific Reports*, *15*(1), 32818.