Machine learning algorithms for predicting the risk of chronic kidney disease in type 1 diabetes patients: a retrospective longitudinal study

Abstract

This research introduces a novel framework designed to accurately predict heart disease while prioritizing stringent patient data privacy. A collaborative system involving multiple entities is established to facilitate efficient data collection, secure management, preprocessing, and analysis of medical data. To safeguard sensitive patient information, a robust encryption scheme is implemented to protect data throughout its lifecycle. A comprehensive ensemble of machine learning algorithms, including Naïve Bayes, Decision Trees, Random Forests, K-Nearest Neighbors, Support Vector Machines, and Long Short-Term Memory networks, is rigorously applied to the preprocessed data. To optimize predictive performance, dimensionality reduction techniques such as Principal Component Analysis and anomaly detection using Isolation Forest are incorporated. Rigorous hyperparameter tuning is performed to fine-tune the models and achieve optimal results. Empirical evaluation demonstrates the superior performance of the proposed framework, with the Random Forest algorithm achieving the highest mean F1-score of 91.02%. This robust performance is attributed to the effective combination of feature engineering, model selection, and hyperparameter optimization. By prioritizing data privacy and leveraging advanced machine learning techniques, this research offers a significant contribution to the field of heart disease prediction.

Keywords Heart disease prediction · Data privacy · Machine learning · Principal Component Analysis · Isolation Forest · Long short-term memory

# Introduction

Heart disease remains a leading cause of global mortality and a significant public health challenge, affecting approximately 26 million individuals worldwide [[1]](#_References), with prevalence expected to rise due to aging populations [[2]](#_References). The increasing burden of cardiovascular disease (CVD) on healthcare systems is evident in its contribution to 17% of overall national health expenditures in the United States alone [[3]](#_References). Despite advancements in diagnostic tools and medical management, limited access to state-of-the-art equipment, especially in rural areas, and reliance on physician intuition continue to impede early detection and effective treatment [[4]](#_References). Recent advancements in machine learning (ML) have demonstrated transformative potential in improving disease diagnosis and predictive accuracy by learning patterns from high-quality datasets [[5, 6]](#_References). This study proposes a robust ML-based system for heart disease prediction, aiming to bridge the diagnostic gap, enhance patient safety, and reduce healthcare costs.

Existing diagnostic methods often fall short in terms of early detection due to limitations in accuracy, efficiency, and interpretability. Traditional models may struggle to capture complex interactions between features and temporal dependencies within the data, leading to suboptimal predictive performance. Additionally, concerns around data privacy hinder the development of robust models.

To address these challenges, this study proposes a novel hybrid approach that combines the strengths of ensemble methods, dimensionality reduction, anomaly detection, and deep learning. The integration of Random Forest Classifier (RF), Principal Component Analysis (PCA), Isolation Forest (iForest), and Long Short-Term Memory (LSTM) contributes to a more robust and accurate predictive model. The LSTM component is particularly crucial for capturing potential temporal patterns in patient data, which traditional ML models might overlook. Furthermore, our emphasis on data security, including encryption and privacy measures, ensures patient confidentiality while enabling effective model development.

This hybrid approach distinguishes itself from existing methods by simultaneously considering multiple feature perspectives (PCA and iForest), capturing temporal dynamics (LSTM), and leveraging the ensemble power of RF. Additionally, data security is reinforced through the use of ElGamal encryption, which ensures that intermediate parties cannot alter the data while it remains encrypted during the application of machine learning techniques. This comprehensive approach is expected to enhance predictive accuracy, provide valuable insights into the underlying patterns of heart disease, and improve patient outcomes by enabling earlier detection and intervention. Such advancements could lead to reduced morbidity and mortality rates associated with heart disease, as well as decreased healthcare costs. Moreover, the proposed model’s emphasis on data security and privacy aligns with emerging healthcare regulations and patient expectations, fostering trust in healthcare systems. Ultimately, the successful implementation of this model has the potential to significantly impact public health by reducing the global burden of heart disease.

The paper systematically evaluates the performance of various ML models and their combinations using a comprehensive set of metrics, including accuracy, precision, recall, F1-score, and Area Under the Curve (AUC). The impact of different training set sizes on model performance is also investigated to assess the robustness of the proposed approach.

# Literature Review

Heart disease prediction has been a focal point of numerous studies, with researchers exploring various ML techniques and feature selection strategies to enhance diagnostic accuracy. One investigation explored the optimization of feature selection and classifier performance in heart disease diagnosis [[7]](#_References). The study compared established techniques (Relief, Minimum Redundancy and Maximum Relevance (MRMR), Least Absolute Shrinkage and Selection Operator (LASSO), LLBFS) with a novel FCMIM algorithm, evaluating their influence on various classifiers, including Logistic Regression (LR), K-Nearest Neighbors (KNN), Artificial Neural Networks (ANN), Support Vector Machines (SVM), Naïve Bayes Classifier (NB), and Decision Tree (DT). The findings revealed that the combination of FCMIM and SVM achieved the highest accuracy of 92.37% using Leave-One-Out Cross-Validation. While Deep Neural Networks (DNN) were explored, limited data hindered their performance.

Another study introduced an HRFLM model combining RF and Linear Model for improved heart disease classification [[8]](#_References). Here, the experimental results showcased significantly higher accuracy compared to traditional methods, underscoring the effectiveness of this approach.

Another investigation aimed to achieve accurate prediction not only for the presence or absence of CVD but also for its severity levels [[9]](#_References). This research employed ML models like SVM, KNN, LR, Stochastic Gradient Descent (SGD), and tree-based ensembles to address the challenge of imbalanced data using Synthetic Minority Oversampling Technique (SMOTE) and hyperparameter optimization with HPO. Notably, tree-based ensemble models outperformed the others, achieving impressive accuracies of 99.2% and 98.52% for CVD presence/absence and 95.73% for severity level prediction.

Further research compared six models, including Extreme Gradient Boosting (XGB), Bagging, RF, DT, KNN, and NB, on a dataset of over 300,000 cases [[10]](#_References). XGB emerged as the top performer with an accuracy of 91.30% and AUC of 0.83. Feature selection using sequential backward selection and KNN was also investigated, achieving a classification accuracy of 90% with six selected features [[11]](#_References).

Subsequent studies compared LR, KNN, SVM, and XGB classifiers, optimized with Grid Search Cross-Validation (GridSearchCV) [[12, 13, 14]](#_References). RF consistently demonstrated superior performance, with accuracies reaching 99% and 98.53%. A more recent study introduced ensemble methods and a novel Quine McCluskey Binary Classifier (QMBC) classifier, along with feature selection techniques like Chi-Square and Analysis of Variance (ANOVA), further improving prediction accuracy [[15]](#_References).

While these studies provide valuable insights, this research distinguishes itself by introducing a novel framework prioritizing data security and privacy, encompassing a complete system lifecycle, and employing rigorous evaluation methods.

# Preliminaries

## ElGamal Cryptosystem

The ElGamal cryptosystem is a public-key encryption scheme based on the discrete logarithm problem. Given a cyclic group G of prime order p with generator g, a private key a is chosen randomly from {1, ..., p−1}. The public key is ga mod p. To encrypt a message m ∈ G, a random integer k is chosen, and the ciphertext is (c1​, c2​) = (gk mod p, m ∗ (ga)k mod p). Decryption is done by computing c2 ​/ (c1a​) mod p [[16]](#_References).

## Naïve Bayes Classifier (NB)

NB is a simple probabilistic classifier that assumes feature independence given the class label. Despite this assumption, it often performs well due to its focus on finding the class with the highest probability. Recent research suggests that NB can be optimal even when feature independence is violated, as long as dependencies balance out across classes [[17]](#_References).

## Decision Tree (DT)

A DT is a model used for classification and regression tasks that makes decisions by splitting data into subsets based on feature values. Each internal node represents a decision point, where a specific feature is selected to split the data. The terminal leaves represent the final decision or output. The tree structure allows for easy interpretation and visualization, but it can become complex with very large datasets [[18]](#_References).

## Random Forest Classifier (RF)

RF is an ensemble method that builds multiple DTs, each trained on a random subset of features and samples. During classification, each tree votes for a class, and the majority vote determines the final output. This approach leverages bagging, where each tree is trained on a bootstrap sample, and uses the Gini Index to evaluate the impurity of nodes. By aggregating the predictions from many trees, RF improves accuracy and reduces overfitting compared to single DTs [[19]](#_References).

## K-Nearest Neighbors (KNN)

The KNN algorithm is a simple yet effective supervised learning technique for both classification and regression tasks. Given a query point, KNN finds the k closest data points in the training set and assigns the query point to the most frequent class among those neighbors (for classification) or the mean of the neighbor values (for regression) [[20]](#_References).

## Support Vector Machine (SVM)

SVM is a robust supervised learning technique for classification tasks. They work by identifying the optimal hyperplane that maximizes the margin between different classes. Data points nearest to the hyperplane, known as support vectors, significantly influence its placement. To handle complex, non-linearly separable datasets, SVMs employ the kernel trick. This technique maps the data into a higher-dimensional space, where it becomes linearly separable [[21]](#_References). SVMs commonly employ kernel functions such as linear, polynomial, RBF, and sigmoid.

## Principal Component Analysis (PCA)

PCA is a dimensionality reduction technique that identifies patterns in data and transforms it into a new coordinate system. Key steps include standardization, covariance matrix calculation, eigenvalue decomposition, principal component selection, and data projection. PCA is used for various purposes like dimensionality reduction, feature extraction, noise reduction, visualization, and outlier detection [[22]](#_References).

## Isolation Forest (iForest)

iForest is an unsupervised anomaly detection method that isolates anomalies using a forest of isolation trees. Anomalies are identified based on their shorter path lengths to isolation compared to normal instances. iForest offers advantages like linear time complexity, handling high-dimensional data, and robustness to noise and outliers. It has shown superior performance in accuracy and efficiency compared to other methods [[23]](#_References).

## Long Short-Term Memory (LSTM)

LSTM networks are a type of recurrent neural network (RNN) designed to address the vanishing gradient problem, which hinders traditional RNNs from learning long-term dependencies in sequential data. LSTMs achieve this by employing a unique cell structure with input, output, and forget gates. These gates regulate the flow of information into and out of the cell, enabling the network to store and retrieve information over extended time periods [[24]](#_References). LSTMs are effective for tasks like natural language processing and speech recognition, where understanding context is crucial.

## Train-Test Split

The train-test split is a method for evaluating ML models by dividing data into a training set (to fit the model) and a test set (to assess performance). Common split ratios are 70:30 or 80:20. Choosing the right ratio is essential, as an imbalanced split can skew performance metrics by underutilizing data or providing insufficient test data [[25]](#_References).

## K-Fold Cross-Validation

K-fold cross-validation is a technique for assessing a model’s performance by dividing the dataset into k equally sized folds, using k−1 folds for training and the remaining fold for testing, and rotating this process across all folds. This ensures that every data point is used for both training and validation [[26]](#_References).

# System Model and Functional Roles

This section provides a detailed architectural blueprint of the system, as shown in Figure 2. It defines the system's components, their interactions, and respective functions. Additionally, it outlines the sequential flow of activities within the system.

## Roles of each entity

The proposed system comprises six primary entities: Client Company (CC), Cryptographer (CR), Service Provider (SP), System Developer (SD), System (SYS) and System User (SU), as illustrated in Figure 1.

*CC*The CC oversees the entire process, starting by gathering raw datasets related to heart disease from various hospitals and combining them into a unified dataset. It generates public-private key pairs, shares the public keys with the CR for data encryption, and securely stores the private keys. Once the encrypted dataset is received from the CR, the CC validates its integrity by decrypting it with the private keys and verifying the data. If the dataset is deemed valid, the CC forwards it to the SP for model training. After receiving the trained models, test datasets, preprocessing scalers, and F1-scores from the SP, the CC tests the models on its own encrypted dataset and compares the F1-scores with those reported by the SP. If the models pass validation, the CC develops the SYS using the best-performing model and preprocessing scaler. This SYS is then deployed for SUs to use.

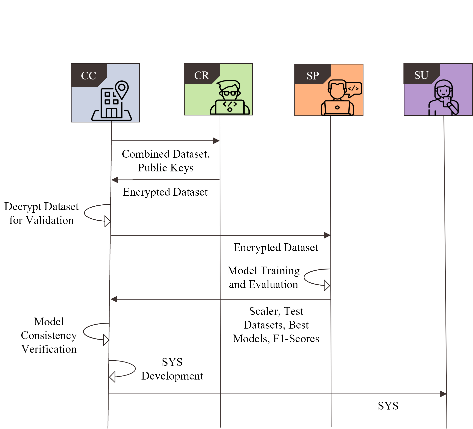


Fig. 1 Interactions between entities

*CR* The CR receives the unified dataset from CC and encrypts it using the public keys shared by CC, ensuring the data remains confidential. Once encrypted, the dataset is sent back to CC.

|  |
| --- |
|  |
| Fig. 2 Proposed Heart Disease Prediction System |

*SP* The SP preprocesses the encrypted dataset received from CC and trains several ML models. After evaluating these models using metrics like F1-score, the SP selects the best-performing model. The trained models, test datasets, preprocessing scalers, and F1-scores for the selected model are then shared with CC.

*SYS* The SYS is the final application developed by CC, using the trained model and preprocessing scaler provided by SP. It takes input data from SUs, encrypts it with the public keys shared by CC, preprocesses the data, and then makes predictions based on the trained model.

*SU* The SUs interact with the SYS to check whether they have heart disease.

## Roles of each entity

Table 1 Notation Used

|  |  |
| --- | --- |
| *Symbol* | *Description* |
| *ωk* | Raw dataset from the k-th Kaggle repository, where k ∈ {1, 2, ..., 7} |
| *ω* | ω1, ω2, …, ω7 |
| *Φ* | Dataset features |
| *d* | No. of crucial features |
| *r* | No. of medical records |
| *ξ* | Crucial features (ξ1,ξ2, …, ξd) |
| *λ* | Label value (Categorial) |
| *i ∈ d* | i-th column |
| *j ∈ r* | j-th row |
| *Δ* | Combined Dataset (Δ1, Δ2, …, Δd, λ) |
| *∅* | Missing value |
| *κ* | ​Public keys (pi, gi, gai) |
| *ρ* | Private keys (ai) |
| *E(κi, ξi)* | ElGamal encryption function |
| *λ'* | Encoded label value |
| *Θ* | Encrypted dataset (Θ1­,Θ2,​ …, Θd, λ’) |
| *D(ρi, Θi)* | ElGamal decryption function |
| *Υ* | Decrypted dataset (Υ1­,Υ2,​ …, Υd, λ’’) |
| *θ* | Standardized Dataset (θ1­,θ2,​ …, θd, λ’) |
| *η1, η2, η3, η4 η5* | ML models: NB, DT, RF, KNN, SVC |
| *η*α | Best performing model |
| *ηβ* | Best configuration of ηα |
| *Π1, Π2, Π3,Π4, Π5,Π6,Π7,Π8* | Confusion Matrix, Accuracy, Specificity, Recall, Positive Predictive Value (PPV), Negative Predictive Value (NPV), F1-score, AUC |
| *τ* | Ratio of train-test split (0.1:0.9 to 0.9:0.1) |
| *δ* | Preprocessing Transformer |

## Individual stages

The proposed framework consists of eight distinct stages. These are: (1) Data Collection, (2) Data Encryption, (3) Data Preprocessing, (4) Model Training, (5) Model Optimization, (6) Final Model selection, (7) Model verification and (8) System Development. These stages are detailed below, with notations referenced from Table 1:

4.3.1 Data Collection

The CC is responsible for collecting datasets from various hospitals for heart disease prediction. For this research, data was gathered from seven publicly available Kaggle repositories [[26-32]](#_References). The datasets, denoted as ω1​, ω2, ..., ω7, contain several inconsistencies. These include irrelevant features that aren't necessary for prediction. The features consist of both numerical and categorical values, and some of them are encoded differently across the ω.

To ensure consistency, CC pre-processes the data by retaining only the relevant features (ξ, where ξ ⊂ Φ) from all available features (Φ), while discarding the irrelevant ones. CC manually transforms all the values across the datasets to ensure a uniform representation, standardizing the types of data within the entire set. To ensure consistency, CC reviews the Kaggle dataset descriptions to identify the data type for each ξi and λ in ω. A standard data type is then chosen for all features and label, and the values in each dataset are converted to this standardized format, ensuring uniformity in attribute names and value types across all datasets.

Once pre-processed, ω are combined into a single, unified dataset, Δ = . CC also generates d pairs of public and private keys (κi and ρi) for d attributes following the ElGamal cryptosystem as described in Section 3.1. The Δ, along with the κ, is then transferred to CR to make the dataset secure.

|  |
| --- |
| Algorithm 1Key Generation (According to Section 3.1) |
| 1. **function** *gen\_key(p):* 2. *a ←* **random int** *[10^20, p]* 3. **while** *gcd(p, a) ≠ 1* **do** 4. *a ←* **random int** *[10^20, p]* 5. **end while** 6. **return** *a* 7. **end function** 8. **function** *generate\_keys():* 9. *p ←* **random int** *[10^20, 10^50]* 10. *g ←* **random int** *[2, p]* 11. *κ, ρ ← [], []* 12. **for** *i* **from** *0* **to** *d* **do** 13. *κ.append(g^gen\_key(p) mod p)* 14. *ρ.append(gen\_key(p))* 15. **end for** 16. **return** *κ, ρ, p, g* 17. **end function** 18. *κ, ρ, p, g ← generate\_keys()* |

Δ has following attributes as detailed in Table 2:

Table 2 Heart Disease Prediction Attributes

|  |  |  |
| --- | --- | --- |
| *Attribute* | *Unit* | *Type of Data* |
| Age | Numeric | In Years |
| Sex | Nominal | 1. Male 2. Female |
| Chest Pain Type | Nominal | 1. Typical Angina 2. Atypical Angina 3. Non-anginal Pain 4. Asymptomatic |
| Resting Blood Pressure | Numeric | 94-200 (mm HG) |
| Cholesterol | Numeric | 126-564 (mg/dl) |
| Fasting Blood Sugar | Binary | Yes / No > 120 mg/dl |
| Resting ECG | Nominal | 1. Normal 2. Abnormal ST-T Wave 3. Left Ventricular Hypertrophy |
| Maximum Heart Rate | Numeric | 71-202 |
| Exercise Angina | Binary | Yes / No |
| Oldpeak | Numeric | 0 - 6.2 |
| ST Slope | Nominal | 1. Upsloping 2. Flat 3. Downsloping |

Figure 3 provides a visual representation of the distribution of key attributes within the dataset.

4.3.2 Data Encryption

The CR secures the dataset Δ, which contains d feature vectors (ξ1, ξ2, …, ξd​) and their associated label λ, by applying the encryption function E(κi, ξi). Each ξi is encrypted as described in Section 3.1, using its corresponding public key κi, provided by CC, resulting in its encrypted form Θi​. For λ, CR employs a straightforward binary encoding scheme, converting 'Yes' to 1 and 'No' to 0, producing λ′. The fully encrypted dataset Θ is then sent to CC for validation.

|  |
| --- |
| Algorithm 2 Data Encryption (According to Section 3.1) |
| 1. **function** *enc(m, p, h, g):* 2. *k ← gen\_key(p)* 3. *s ← h^k mod p* 4. *p ← g^k mod p* 5. **return** *[s \* v* **for** *v* **in** *m], p* 6. **end function** 7. **for** *i* **from**0**to** *d* **do** 8. *encoded\_vals ← encode(Δi)* 9. **for** *v* **in** *encoded\_vals* **do** 10. **if** *v* **is valid****do** 11. *enc, p ← enc(v, q, κi, g)* 12. **end if** 13. *Θi.append(enc)* 14. *pi.append(p)* 15. **end for** 16. *Θ.append(Θi)* 17. *p\_list.append(pi)* 18. **end for** |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| **(a)** | **(b)** | **(c)** | **(d)** | **(e)** | **(f)** |
|  |  |  |  |  |  |
| **(g)** | **(h)** | **(i)** | **(j)** | **(k)** |  |
| Fig. 3 Distribution of Heart Disease Prediction Attributes | | | | | |

4.3.3 Data Decryption and Verification

The CC uses the private keys ρi and the decryption function D(ρi, Θi) to decrypt Θi, obtaining the decrypted features Υi. The CC also decodes the labels λ′, converting 1 back to ‘Yes’ and 0 to ‘No,’ to produce λ′′. The CC then compares Υi​ and λ′′ with the original Δi and λ to ensure no data was altered during encryption or transmission by the CR. If all Δi and λ values match their corresponding Υi and λ′′ values, the CC forwards Θ to the SP for model training.

|  |
| --- |
| Algorithm 3 Data Decryption and Verification |
| 1. **function** *dec(enc, p, x, q):* 2. *h ← p^x mod q* 3. **return** *[v / h* **for** *v* **in** *enc]* 4. **for** *i* **from** 0 **to** *d* **do** 5. **for each** *(enc, p)* **in** **zip**(*Θi*, p\_listi) **do** 6. **if** *enc* **and** *p* **are valid do** 7. *d ← dec(enc, p, ρi, q)[0]* 8. **end if** 9. *Υi.append(d)* 10. **end for** 11. *labels ← factorize(Δi)* 12. *decoded\_vals ← [labels[v]* **if** *v* **is valid****else****None****for** *v* **in** *Υi]* 13. **for each** *(dec, org)* **in zip***(decoded\_vals, Δi)* **do** 14. **if** *dec ≠ org* **do** 15. **count** *mismatch* 16. **end if** 17. **end for** 18. **end for** |

4.3.4 Data Preprocessing

The SP addresses missing values ∅ in Θ by imputing the mean value for each Θi​. Missing values in Θij​ are replaced with the mean of all non-missing values in Θi​, calculated as:

After imputation, Θij​ is standardized using the StandardScaler technique, denoted as δ. Standardization ensures all features have a mean of zero and a variance of one, transforming Θij​ into θij​:

where,

and,

*Standard Deviation,*

The entire process, handling missing values through mean imputation and scaling the data via standardization, is denoted as δ for reference.

4.3.5 Model Training

The SP partitions θ into training (θtrain) and testing (θtest) subsets. It then evaluates a selection of ML models, including NB (η1), DT(η2), RF(η3), KNN(η4), and SVC(η5), chosen for their ability to handle categorical data without requiring decrypted features. For each model, the SP computes various performance metrics: Confusion Matrix (Π1), Accuracy(Π2), Specificity(Π3), Recall(Π4), PPV(Π5), NPV(Π6), F1-score(Π7) and AUC(Π8).

Additionally, the SP performs hyperparameter optimization to maximize the Π7 for each model, as the rationale for selecting Π7​ is explained in Section 4. To further refine the analysis, the SP explores the impact of different train-test split ratios (τ) on model performance, evaluating Π7​ for τ values ranging from 0.1:0.9 to 0.9:0.1 across all models. Based on the highest Π7​ values, the SP selects the best-performing model (ηα​) for further evaluation and potential hyperparameter tuning.

4.3.6 Model Optimization

The SP conducts a more in-depth evaluation on ηα, the model with the optimal hyperparameters that yielded the highest Π7. SP uses 10-fold cross-validation, dividing the θ into 10 parts, where each part is used once as a validation set while the others form the training set, ensuring robust and unbiased evaluation. Additionally, SP utilizes three advanced techniques for analysis: PCA, iForest, and LSTM.

The hyperparameters optimized for the models are:

1. RF: The number of trees (n\_estimators), maximum tree depth (max\_depth), minimum samples required to split nodes (min\_samples\_split), and split criterion (Gini).
2. PCA: The number of principal components to retain (n\_components) to balance dimensionality reduction and variance.
3. iForest: The number of trees (n\_estimators), contamination rate (the proportion of outliers expected in the data), and sample size (max\_samples) for anomaly detection.
4. LSTM: The number of LSTM units, learning rate, dropout rate, number of epochs, and batch size for sequential tasks.

4.3.7 Final Model Selection

The SP conducts a comprehensive evaluation of four configurations of the best-performing model ηα: the base model (ηα1), ηα with PCA (ηα2), ηα with iForest (ηα3), and ηα with LSTM (ηα4). For each configuration, SP applies 10-fold cross-validation, generating 10 sets of θtrain and θtest, where each set includes the randomly selected rows along with their corresponding serial numbers. Additionally, 10 trained models (ηαl1, ..., ηαl10, where l ∈ {1, 2, 3, 4}), and 10 corresponding Π7 scores. The average Π7 score across the 10 folds, denoted as Π7’, is calculated for each configuration.

The SP selects the configuration with the highest average Π7’ and identifies it as ηβ from {ηα1, ηα2, ηα3, ηα4}. All 10 models of ηβ, along with their corresponding θtest, Π7, and δ, are then delivered to the CC for further validation.

4.3.8 Model Verification

The CC validates the models ηβ trained by the SP. For this, the CC randomly selects some or all of the θtest and retrieves the corresponding serial numbers from θ. Using these serial numbers, the CC extracts the corresponding rows from its own Θ. This subset of Θ is preprocessed using δ to produce θ and is then trained using the corresponding ηβ.

The CC then calculates the performance metric Π7 and compares it with the Π7 provided by the SP. If the calculated Π7 match exactly with those from the SP, the CC proceeds to develop the SYS.

|  |
| --- |
| Algorithm 4Model Verification |
| 1. 𝛾 *← []* // true labels 2. 𝜋 *← []* // predicted labels 3. **for** *j* **from** 0 **to** *r*: 4. *Θj* ← *δ.transform(Θj)* 5. *𝑝j ← ηβ.predict(Θj)* 6. *𝛾.append(λj')* 7. *𝜋.append(𝑝j[0])* 9. *𝑓1 ← f1\_score(𝛾, 𝜋) \** 100 |

4.3.9 System Development

The CC develops a user-friendly interface for SU interaction. This interface includes input fields for each of the model's d features (ξi), allowing SUs to input their data directly. The system encrypts each ξi using its corresponding public key κi, producing Θi. These encrypted values are then processed using δ to generate θi, which is fed into the ηβ model with the highest Π7 to predict λi'. Once finalized, the CC makes the system publicly available for SUs.

# Security Threats and Solutions

In our proposed heart system, several potential security threats were identified, and corresponding measures were implemented to address these threats. This section outlines the primary threats and the solutions incorporated in our methodology to ensure data integrity and system security.

## Data Confidentiality Breach

*Threat:* Unauthorized access to sensitive medical data could occur during transmission or storage, leading to potential data leaks.

*Solution:* To address this, SD uses the ElGamal encryption function E(.) to encrypt the dataset with unique keys for each feature. This ensures that the data remains confidential even if intercepted.

## Predictable Encryption

*Threat:* Using a single encryption key for all features could create predictable encryption patterns, making the data vulnerable to attacks.

*Solution:* To mitigate this risk, SD generate unique encryption keys for each feature. This prevents identical ciphertexts for identical values, thereby enhancing encryption security.

## Integrity of Model Training Data

*Threat:* Tampering with the training data could result in inaccurate model predictions or biases.

*Solution:* The system incorporates a validation step where the SD verifies the model’s integrity by comparing predictions from encrypted and original datasets. Any inconsistencies prompt further investigation to ensure the model's accuracy and reliability.

## Replay Attacks

*Threat:* Captured encrypted data might be replayed to the system, potentially leading to unauthorized access or manipulation.

*Solution:* To prevent replay attacks, unique encryption keys are generated for each feature, and data is modded by a large prime number. This ensures that each dataset is unique, even if the underlying data is unchanged.

## Replay Attacks

*Threat:* Sensitive information might be inferred through model queries.

*Solution:* To mitigate this risk, models are selected based on their ability to function without requiring decrypted features. This ensures that sensitive data remains encrypted throughout the inference process.

# Results

The performance of machine learning models was evaluated using several key metrics. To comprehensively assess model accuracy, a confusion matrix was employed, offering a tabular representation of the model's predictions compared to the actual ground truth values.

A confusion matrix, as detailed in [[15]](#_References), is a table that summarizes the performance of a classification model by comparing its predicted values with actual values. It categorizes predictions into four groups: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

The following performance metrics were calculated based on the confusion matrix:

*Accuracy* represents the proportion of correct predictions (both positive and negative) out of the total number of instances.

*Accuracy =*

*Recall (Sensitivity)* refers to the percentage of true positive instances that are correctly identified by the model.

*Recall =*

*Specificity* measures the proportion of true negatives correctly identified as such out of all actual negative cases.

*Specificity =*

*Precision or Positive Predictive Value (PPV)* measures the proportion of correctly predicted positive cases out of all instances predicted as positive.

*Precision or PPV =*

*Negative Predictive Value (NPV)* measures the proportion of actual negative cases that are correctly predicted as negative out of all predicted negative cases.

*NPV =*

*F1-Score* combines precision and recall into a single metric by calculating their harmonic mean, balancing both aspects of model performance.

*F1-Score =*

*Area Under the Curve (AUC)* represents the overall performance of a classification model across all possible classification thresholds. A higher AUC value indicates better model discriminative power.

Training ML models on encrypted data is challenging due to the loss of interpretability. Traditional models rely on specific features, which are obscured by encryption. However, the consistent nature of encryption can be exploited to identify relationships within the data.

While challenging, some ML algorithms like NB, DT, RF, SVM, and KNN can be adapted to encrypted data. NB leverages feature independence, DT and RF utilize relative feature ordering, SVM compares relative features, and KNN exploits distance relationships between data points.

Model robustness and generalization were assessed by employing various train-test splits, with the training-to-testing data ratio ranging from 1:9 to 9:1. Models were trained on the training data and evaluated on the corresponding test data using metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. This approach helped identify potential overfitting or underfitting and ensured the reliability and generalization of the models.

|  |
| --- |
|  |
| Fig. 4 Performance Metrics of Classification Algorithms with Varying Train-Test Split |

As the training set size increases, the performance of most models improves, as seen in the upward trends in the graphs. However, some models might be more sensitive to changes in training set size than others.

Overfitting can be observed when a model performs exceptionally well on the training data but poorly on the test data, especially with larger training sets. Underfitting occurs when a model consistently performs poorly, regardless of the training set size.

Figure 4 illustrates how the performance of different machine learning models on test data varies with changes in train-test split. The evaluated models are NB, DT, RF, KNN, and SVM.

RF consistently outperforms other algorithms across all evaluation metrics. It achieves excellent performance in accuracy (average 89.98%), specificity (average 89.31%), recall (average 90.54%), PPV (average 91.55%), NPV (87.98%), F1-score (average 91.04%), and AUC (average 94.63%). These results indicate that RF is highly effective in classifying the dataset.

SVM demonstrates strong performance, particularly in accuracy (79.24%), specificity (71.91%), recall (84.88%), PPV (79.67%), NPV (78.61%), and F1-score (82.18%). While not surpassing RF, SVM provides robust classification results.

DT exhibits acceptable performance across most metrics, with accuracy (85.94%), specificity (85.17%), recall (86.57%), PPV (87.41%), NPV (84.34%), F1-score (86.96%), and AUC (85.87%) falling within the good performance range.

NB and K-NN demonstrate similar performance patterns, with strengths in accuracy, specificity, and recall but weaknesses in predictive values. Both models show room for improvement, especially in terms of positive and negative predictive values.

Increasing training set size generally improves performance for all algorithms, but the rate of improvement diminishes after a certain point. Increasing the train-test split ratio from 0.1 to 0.9 helped identify trends in model behavior and provided evidence of stable generalization across varying dataset sizes. None of the models showed significant signs of overfitting or underfitting, indicating that both the dataset and the models are well-suited for classification tasks.

RF is chosen as the best model due to its consistently superior performance across all metrics, demonstrating its ability to effectively classify the dataset with minimal loss of precision or recall. It consistently achieves the highest accuracy, F1-score, and AUC, indicating its robustness and generalization ability. In contrast, DT, NB, and K-NN perform relatively well in some metrics, but their performance is less consistent and shows weaknesses in predictive values, particularly in PPV and NPV. These limitations make DT, NB, and K-NN less reliable for precise and consistent classification.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 3 Performance Comparison of Classification Models with Varying Training Set Sizes and Hyperparameters (F1-Score in %) | | | | | | | | | | | | | | |
| *Classifiers* | *Hyper-parameters* | | | *Train-Test Ratio* | | | | | | | | | *Mean* | |
| *1:9* | *2:8* | *3:7* | *4:6* | *5:5* | *6:4* | *7:3* | *8:2* | *9:1* |
| *RF* | Data Splitting (train\_test\_split)   * random\_state: 1694 * n\_estimators: 95 | RandomForestClassifier   * random\_state: 1694 |  | 82.69 | 85.31 | 87.81 | 90.71 | 92.51 | 93.75 | 94.42 | 95.24 | 95.82 | 90.92 |
| *RF PCA* | Data Splitting (train\_test\_split)   * random\_state: 983 | PCA   * n\_components: 9 | RandomForestClassifier   * random\_state: 983 * n\_estimators: 91 | 79.24 | 82.58 | 85.18 | 87.06 | 89.03 | 90.86 | 91.56 | 93.01 | 92.75 | 87.92 |
| *RF iForest* | Data Splitting (train\_test\_split)   * random\_state: 1341 | IsolationForest:   * n\_estimators: 26 * contamination: 0.01 * random\_state: 530 * max\_samples: 105 | RandomForestClassifier   * random\_state: 530 * n\_estimators: 63 * criterion: gini | 83.22 | 85.00 | 88.08 | 90.54 | 92.55 | 93.62 | 94.38 | 95.84 | 95.96 | 91.02 |
| *RF LSTM* | Data Splitting (train\_test\_split)   * random\_state: 47 | Optimizer (Adam)   * learning\_rate: ~0.0667 (1/15) | Model Training (model.fit)   * epochs: 107 * batch\_size: 10000 * verbose: 0 | 74.30 | 79.53 | 79.56 | 81.98 | 83.95 | 85.91 | 87.50 | 87.96 | 89.91 | 83.40 |
| Sequential  Dropout   * rate: 0.05   Dense   * units: 1 * activation: sigmoid | Model Compilation (model.compile)   * loss: binary\_crossentropy | RandomForestClassifier   * random\_state: 47 * n\_estimators: 47 |

While all the performance metrics offer valuable insights, the F1-score is prioritized for comparing the models. This is due to its ability to balance precision and recall, both of which are critical in this context. Precision ensures that positive predictions are accurate, while recall ensures that a high proportion of actual positive cases are identified. By optimizing the F1-score, the goal is to develop a model that minimizes both false positives and false negatives.

Although accuracy might seem appealing, it can be misleading in imbalanced datasets, where the majority class dominates the evaluation. PPV focuses on the accuracy of positive predictions, but it might not capture the ability of the model to identify all positive cases. NPV is valuable for assessing the reliability of negative predictions, but it's less critical in our scenario. AUC provides a global measure of model performance, but it doesn't directly address the trade-off between precision and recall. Therefore, the F1-score emerges as the most suitable metric for our comparative analysis.

Table 3 presents a comparative analysis of various RF-based models, including RF, RF with dimensionality reduction (PCA), RF with anomaly detection (Isolation Forest), and RF with LSTM, evaluated across different training set sizes ranging from 10% to 90%.

RF with Isolation Forest demonstrated the highest mean F1-score of 91.02%, outperforming other configurations. RF achieved a mean F1-score of 90.92%, while RF with PCA showed a mean F1-score of 87.92%. RF with LSTM achieved the lowest mean F1-score of 83.40%.

Dimensionality reduction through PCA showed mixed results, with improvements at smaller training set sizes, but diminishing returns for larger datasets. LSTM-based models showed potential but did not outperform RF-based ensemble methods.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Table 4: Comparative Analysis of Proposed System with Existing Methods | | | | | | |
| *Attributes* | [*[7]*](#_References) | [*[8]*](#_References) | [*[9]*](#_References) | [*[12]*](#_References) | [*[15]*](#_References) | *Proposed System* |
| *Dataset Size and Diversity* | Cleveland dataset | Cleveland dataset | Two public datasets | Cleveland, Hungary, Switzerland, Long Beach V, UCI Kaggle | Cleveland, Comprehensive, CVD datasets | Combined hospital data |
| *Data Quality and Preprocessing Techniques* | Missing values removed, SS, Min-Max | Missing values removal, Binary conversion | SMOTE, Splitting, Balancing | Quality implication, Standard preprocessing | Data type conversion, Outliers handling, SMOTE, Under-sampling | Standardization, Scaling, Encoding, PCA, iForest, Data cleaning, Feature selection |
| *Feature Engineering* | Relief, MRMR, LASSO | HRFLM method | Significant features | 14-attribute dataset created | Chi-Square, ANOVA, FS, FE | - |
| *ML Algorithms* | ANN, LR, K-NN, SVM, DT, NB | NB, GLM, LR, DL, DT, RF, GBT, SVM | SVM, SGD, K-NN, ET, XGB, LR | LR, K-NN, SVM, GBC | LR, DT, RF, K-NN, NB, SVM, MLP, QMBC | NB, DT, RF, KNN, SVM |
| *Hyperparameter Tuning* | Leave-one-subject-out | - | Hyperband, SMOTE | GridSearchCV | Minimal, Basic evaluation | Optimized "random\_state" |
| *Evaluation Metrics* | Accuracy, Specificity, Sensitivity, MCC | Accuracy, Sensitivity, Specificity, Precision, F-Measure | Accuracy, F1-Score, MCC | Accuracy, Precision, Recall, F1-Score | Accuracy, Sensitivity, Specificity, Precision, Recall, F1-Score, AUC | Accuracy, Recall, Specificity, Precision, F1-Score, AUC |
| *ML Model, Train-Test Ratio, Evaluation Metric* | SVM+Relief+ LASSO+FCMIM+LLBFS, Accuracy (92.37%) | HRFLM (RF+LM), 7:3, Accuracy (88.7%) | HB+SMOTE+ET, 3:1, F1-Score (95.78%) | XGB+ GridSearchCV, 7:3, Accuracy (99.03%) | Cleveland Dataset: QMBC+Anova+ PCA, 8:2, F1-Score (98.59%) | RF+iForest, 8:2, F1-Score (95.84%) |
| CVD Dataset: QMBC+Chi-Square+PCA, 8:2. F1-Score (99.92%) |
| HD Dataset: QMBC+Anova+ PCA, 8:2. F1-Score (98.42%) |
| *Novelty of Approach* | Fast mutual information | HRFLM hybrid model | Hyperband, balancing | Near-zero prediction error, unique model | Novel QMBC, improved performance | Encryption, integrated techniques |

A comparative analysis of the proposed system with existing methodologies is presented in Table 4. This table systematically contrasts various attributes, including dataset characteristics, preprocessing techniques, feature engineering, ML algorithms, hyperparameter tuning, evaluation metrics, and performance across different approaches.

Our proposed system utilizes a combined hospital dataset, enhancing data diversity and robustness. This contrasts with smaller datasets employed in previous studies [7, 8]. Rigorous preprocessing, including standardization, scaling, and feature selection, differentiates our approach from those using basic methods or techniques like SMOTE [9, 12].

By incorporating a diverse set of ML algorithms and optimizing hyperparameters, our system achieves superior performance metrics, such as accuracy, sensitivity, specificity, and F1-score, compared to existing models like HRFLM [[8]](#_References) and Hyperband [[9]](#_References). The results presented in Table 4 underscore the effectiveness and reliability of our proposed system, positioning it as a significant advancement in the field.

# Conclusion and Future Works

This paper presents a framework for heart disease prediction, prioritizing data privacy and security. It employs a multi-step approach, including data collection, encryption, preprocessing, model training (DT, RF, KNN, SVM), and evaluation. By combining these techniques with dimensionality reduction and anomaly detection, the framework aims to enhance model performance and robustness.

The experimental results demonstrate the effectiveness of RF and its ensemble variants. Notably, the combination of RF with iForest achieved the highest mean F1-score of 91.02%, surpassing other configurations. While DT achieved respectable mean F1-scores (82.45% to 87.23%), their performance was generally lower compared to RF- based models. Dimensionality reduction through PCA yielded mixed results, with improvements observed for smaller training set sizes (up to 5%) but diminishing returns for larger datasets. LSTM-based models showed potential but did not match the performance of ensemble methods (achieving an average 10% lower F1-score).

The proposed system prioritizes data privacy by encrypting sensitive patient information and ensuring that model training is conducted on encrypted data. This approach safeguards patient confidentiality while enabling accurate heart disease prediction.

Future work could explore the integration of advanced deep learning architectures, such as CNNs and RNNs, to potentially improve prediction accuracy. Additionally, investigating the impact of different encryption algorithms and their computational efficiency would be valuable.

By addressing the challenges of data privacy, model accuracy (achieving over 90% F1-score for RF ensembles), and computational efficiency, this research contributes to the development of reliable and secure heart disease prediction systems.

# References

1. Ponikowski P, Anker SD, AlHabib KF, Cowie MR, Force TL, Hu S, Jaarsma T, Krum H, Rastogi V, Rohde LE, Samal UC. Heart failure: preventing disease and death worldwide. ESC heart failure. 2014 Sep;1(1):4-25.
2. Makhlouf A, Boudouane I, Saadia N, Ramdane Cherif A. Ambient assistance service for fall and heart problem detection. Journal of Ambient Intelligence and Humanized Computing. 2019 Apr 12;10:1527-46.
3. Heidenreich PA, Trogdon JG, Khavjou OA, Butler J, Dracup K, Ezekowitz MD, Finkelstein EA, Hong Y, Johnston SC, Khera A, Lloyd-Jones DM. Forecasting the future of cardiovascular disease in the United States: a policy statement from the American Heart Association. Circulation. 2011 Mar 1;123(8):933-44.
4. Al-Shayea QK. Artificial neural networks in medical diagnosis. International Journal of Computer Science Issues. 2011 Mar 1;8(2):150-4.
5. Wang S, Summers RM. Machine learning and radiology. Medical image analysis. 2012 Jul 1;16(5):933-51.
6. Ghwanmeh S, Mohammad A, Al-Ibrahim A. Innovative artificial neural networks-based decision support system for heart diseases diagnosis.
7. Li JP, Haq AU, Din SU, Khan J, Khan A, Saboor A. Heart disease identification method using machine learning classification in e-healthcare. IEEE access. 2020 Jun 9;8:107562-82.
8. Mohan S, Thirumalai C, Srivastava G. Effective heart disease prediction using hybrid machine learning techniques. IEEE access. 2019 Jun 19;7:81542-54.
9. Abdellatif A, Abdellatef H, Kanesan J, Chow CO, Chuah JH, Gheni HM. An effective heart disease detection and severity level classification model using machine learning and hyperparameter optimization methods. ieee access. 2022 Jul 18;10:79974-85.
10. Das RC, Das MC, Hossain MA, Rahman MA, Hossen MH, Hasan R. Heart disease detection using ml. In2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC) 2023 Mar 8 (pp. 0983-0987). IEEE.
11. Haq AU, Li J, Memon MH, Memon MH, Khan J, Marium SM. Heart disease prediction system using model of machine learning and sequential backward selection algorithm for features selection. In2019 IEEE 5th International Conference for Convergence in Technology (I2CT) 2019 Mar 29 (pp. 1-4). IEEE.
12. Ahmad GN, Fatima H, Ullah S, Saidi AS. Efficient medical diagnosis of human heart diseases using machine learning techniques with and without GridSearchCV. IEEE Access. 2022 Apr 8;10:80151-73.
13. Sharma V, Yadav S, Gupta M. Heart disease prediction using machine learning techniques. In2020 2nd international conference on advances in computing, communication control and networking (ICACCCN) 2020 Dec 18 (pp. 177-181). IEEE.
14. Patidar S, Jain A, Gupta A. Comparative analysis of machine learning algorithms for heart disease predictions. In2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS) 2022 May 25 (pp. 1340-1344). IEEE.
15. Kapila R, Ragunathan T, Saleti S, Lakshmi TJ, Ahmad MW. Heart disease prediction using novel quine McCluskey binary classifier (QMBC). IEEE Access. 2023 Jun 26;11:64324-47.
16. Haraty RA, Otrok H, El-Kassar AN. A comparitive study of elgamal based cryptographic algorithms. InInternational Conference on Enterprise Information Systems 2004 Apr 14 (Vol. 4, pp. 79-84). SCITEPRESS.
17. Zhang H. The optimality of naive Bayes. Aa. 2004 Jan 1;1(2):3.
18. Hall LO, Chawla N, Bowyer KW. Decision tree learning on very large data sets. InSMC'98 Conference Proceedings. 1998 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No. 98CH36218) 1998 Oct 14 (Vol. 3, pp. 2579-2584). IEEE.
19. Pal M. Random forest classifier for remote sensing classification. International journal of remote sensing. 2005 Jan 1;26(1):217-22.
20. Zhang S, Li X, Zong M, Zhu X, Wang R. Efficient kNN classification with different numbers of nearest neighbors. IEEE transactions on neural networks and learning systems. 2017 Apr 12;29(5):1774-85.
21. Jakkula V. Tutorial on support vector machine (svm). School of EECS, Washington State University. 2006;37(2.5):3.
22. Wold S, Esbensen K, Geladi P. Principal component analysis. Chemometrics and intelligent laboratory systems. 1987 Aug 1;2(1-3):37-52.
23. Liu FT, Ting KM, Zhou ZH. Isolation forest. In2008 eighth ieee international conference on data mining 2008 Dec 15 (pp. 413-422). IEEE.
24. Schmidhuber J, Hochreiter S. Long short-term memory. Neural Comput. 1997 Nov 15;9(8):1735-80.
25. Muraina I. Ideal dataset splitting ratios in machine learning algorithms: general concerns for data scientists and data analysts. In7th international Mardin Artuklu scientific research conference 2022 May 1 (pp. 496-504).
26. Anguita D, Ghelardoni L, Ghio A, Oneto L, Ridella S. The'K'in K-fold Cross Validation. InESANN 2012 Apr 25 (Vol. 102, pp. 441-446).