**CSE 4000: Thesis/ Project**

**Protecting Patient Privacy While Improving Heart Disease Diagnosis: A Machine Learning Approach**

By

**Israt Tasnim Esha**

Roll: 1907090



##### Department of Computer Science and Engineering

##### Khulna University of Engineering & Technology

##### Khulna 9203, Bangladesh

**October, 2024**

**CSE 4000: Thesis/ Project**

**Protecting Patient Privacy While Improving Heart Disease Diagnosis: A Machine Learning Approach**

By

**Israt Tasnim Esha**

Roll: 1907090

A thesis submitted in partial fulfillment of the requirements for the degree of

“Bachelor of Science in Computer Science & Engineering”

**Supervisor:**

Dola Das

Assistant Professor

Dept. of Computer Science and Engineering Signature

Khulna University of Engineering & Technology

##### Department of Computer Science and Engineering

##### Khulna University of Engineering & Technology

##### Khulna 9203, Bangladesh

October, 2024

**Acknowledgment**

All the praise to the almighty Allah, whose blessing and mercy has guided us to work fairly. After that, I humbly acknowledge the valuable suggestions, advice, guidance and sincere cooperation of Dola Das, Assistant Professor, department of Computer Science and Engineering, Khulna University of Engineering & Technology, under whose supervision this work was carried out. Her intellectual advice, encouragement and guidance have made me feel confident. Last but not least, I wish to thank my parents, relatives, and friends for their constant supports.

##### Author

**Abstract**

Heart disease remains one of the most prevalent causes of mortality, making early and accurate diagnosis critical. This paper proposes a comprehensive heart disease prediction system that integrates machine learning with blockchain technology to enhance accuracy and secure data management. The system involves seven key entities: the Client Company, System Developer, Service Provider, System User, System, Miners, and Blockchain. The CC collects, processes, and securely hands over heart disease-related datasets to the SD, who encrypts them before sharing with the SP. The SP pre-processes the encrypted data and trains multiple ML models, selecting the best-performing one for deployment. The system allows users to input personal data to receive heart disease predictions, explains feature significance, and offers possible outcomes even with missing data. Additionally, users can securely store their medical history on the blockchain, which ensures data immutability and integrity through the mining process. By leveraging the strengths of ML for predictive accuracy and blockchain for data security, this system provides a reliable and secure tool for both patients and healthcare providers.

**Contents**

|  |  |
| --- | --- |
|  | **Page** |
| Acknowledgement | ii |
| Abstract | iii |
| Contents | iv |
| List of Tables | v |
| List of Figures | vi |

|  |  |  |
| --- | --- | --- |
| **Chapter I** | **Introduction** | 1 |
|  | 1.1 Background | 1 |
|  | 1.1.1 Challenges in Existing Solutions  1.1.2 How Our Proposed System Overcomes Existing Challenges | 1  2 |
|  | 1.2 Objectives | 2 |
|  | 1.3 Scopes | 3 |
|  | 1.4 Unfamiliarity of the Problem | 4 |
|  | 1.5 Project planning | 4 |
|  | 1.6 Applications of the Work | 5 |
|  | 1.7 Organization of the Report | 6 |
| **Chapter II** | **Literature Review** | 7 |
|  | 2.1 Literature Review | 7 |
|  | 2.2 Discussion of Research Gap Solution | 8 |
| **Chapter III** | **Methodology** | 10 |
|  | 3.1 Proposed Methodology | 10 |
|  | 3.2 System Architecture | 11 |
| **Chapter IV** | **Implementation, Result and Discussions** | 14 |
|  | 4.1 Experimental Setup | 14 |
|  | 4.2 Evaluation Metrics | 15 |
|  | 4.3 Dataset | 16 |
|  | 4.4 Implementation and Result | 16 |
|  | 4.4.1 Implementation  4.4.2 Result | 16  21 |
|  | 4.5 Objective Achieved | 23 |
|  | 4.6 Ethical Considerations | 24 |
|  | 4.7 Socio-Economic Impact And Sustainability | 24 |
|  | 4.7.1 Socio-Economic Impact  4.7.2 Sustainability | 24  25 |
|  | 4.8 Financial Analyses and Budget | 25 |
| **Chapter V** | **Conclusions** | 26 |
|  | 5.1 Summary | 26 |
|  | 5.2 Limitations | 26 |
|  | 5.3 Recommendations and Future Works | 26 |
|  | **References** | 28 |

**List of Tables**

|  |  |  |
| --- | --- | --- |
| **Table No.** | **Description** | **Page** |
| 2.1 | Contributions and Limitations in Existing Solutions | 8 |
| 4.1 | Heart Disease Prediction Attributes | 17 |
| 4.2 | Financial Analyses and Budget | 25 |

**List of Figures**

|  |  |  |
| --- | --- | --- |
| **Figure No.** | **Description** | **Page** |
| 1.1 | Project Timeline Overview for Thesis Work | 5 |
| 3.1 | Interactions Between Seven Entities | 10 |
| 3.2 | Proposed System Architecture (Part-1) | 11 |
| 3.3 | Proposed System Architecture (Part-2) | 12 |
| 3.4 | Proposed System Architecture (Part-3) | 13 |
| 4.1 | System Architecture | 17 |
| 4.2 | Distribution of Features by Heart Disease Status | 18 |
| 4.3 | Performance Metrics of Classification Algorithms | 22 |

**CHAPTER I**

**Introduction**

**1.1 Background**

Heart disease, a leading cause of death worldwide, poses a significant challenge to public health. The World Health Organization (WHO) estimates that over 23.6 million people will die from cardiovascular disease (CVD) by 2030 [1]. Early detection and accurate diagnosis are crucial for effective treatment and improved patient outcomes. However, existing diagnostic methods often fall short in terms of accuracy, efficiency, and interpretability.

Machine learning (ML) has emerged as a promising tool for addressing these challenges. By analyzing large datasets of patient information, ML algorithms can identify patterns and predict the risk of heart disease. Previous studies have explored various ML techniques and feature selection strategies to enhance diagnostic accuracy. For example, ensemble methods, such as random forests, have shown promising results in heart disease prediction due to their ability to combine multiple models and reduce overfitting [2].

However, existing ML approaches often face limitations in terms of data privacy and interpretability. Concerns about data security hinder the development of robust models, and traditional ML models may struggle to capture complex interactions between features and temporal dependencies within the data.

**1.1.1 Challenges in Existing Solutions**

* **Data Privacy:** Traditional ML models often require sharing sensitive patient data, raising concerns about data privacy and security. This can hinder the development and deployment of robust models, especially in healthcare settings where patient confidentiality is paramount.
* **Computational Efficiency:** Some ML models, particularly deep learning models, can be computationally expensive to train and deploy, making them impractical for real-time applications in resource-constrained environments.
* **Interpretability:** Many ML models are considered black-box models, making it difficult to understand how they arrive at their predictions. This lack of interpretability can limit their clinical utility and hinder trust in the model's outputs.
* **Temporal Dynamics:** Heart disease is a complex condition influenced by a variety of factors that may change over time. Existing ML models may struggle to capture these temporal dependencies, leading to suboptimal predictive performance.

**1.1.2 How Our Proposed System Overcomes Existing Challenges**

Our proposed hybrid approach addresses the aforementioned challenges by incorporating the following key features:

* **Data Security and Privacy:** We employ encryption techniques to protect patient data during transmission and storage, ensuring confidentiality and compliance with data privacy regulations. Additionally, we leverage differential privacy techniques to further enhance data privacy while enabling model development.
* **Computational Efficiency:** Our proposed system combines ensemble methods, dimensionality reduction, and anomaly detection techniques, which can be computationally more efficient than complex deep learning models. This allows for faster training and deployment, making the system suitable for real-time applications.
* **Interpretability:** By incorporating interpretable ML models and feature selection techniques, we aim to provide insights into the factors that contribute to heart disease risk. This can help clinicians understand the model's predictions and make informed decisions.
* **Temporal Dynamics:** The LSTM component of our model is specifically designed to capture temporal dependencies in patient data, allowing us to better account for the dynamic nature of heart disease.
* **Blockchain Integration:** Our system incorporates blockchain technology to ensure data integrity and provenance. This can help prevent data tampering and provide a secure and transparent record of patient data.

By addressing these challenges, our proposed system aims to provide a more robust, accurate, and interpretable solution for heart disease prediction, ultimately improving patient outcomes and reducing the global burden of this disease.

**1.2 Objectives**

The primary objectives of this research are as follows:

* **Develop a robust and accurate heart disease prediction model:** We aim to create a model that can accurately predict the presence or absence of heart disease, considering the complex interplay of various factors and temporal dependencies.
* **Prioritize data security and privacy:** We will ensure that patient data is handled securely and confidentially throughout the entire process, adhering to strict data privacy regulations.
* **Leverage the power of ensemble methods, dimensionality reduction, anomaly detection, and deep learning:** By combining these techniques, we seek to enhance the model's predictive performance and interpretability.
* **Explore the impact of different training set sizes on model performance:** We will investigate how the size of the training dataset affects the model's accuracy and generalization capabilities.
* **Integrate blockchain technology for secure data storage and management:** We will utilize blockchain to ensure data integrity, prevent tampering, and provide a transparent and secure record of patient health information.

**1.3 Scope**

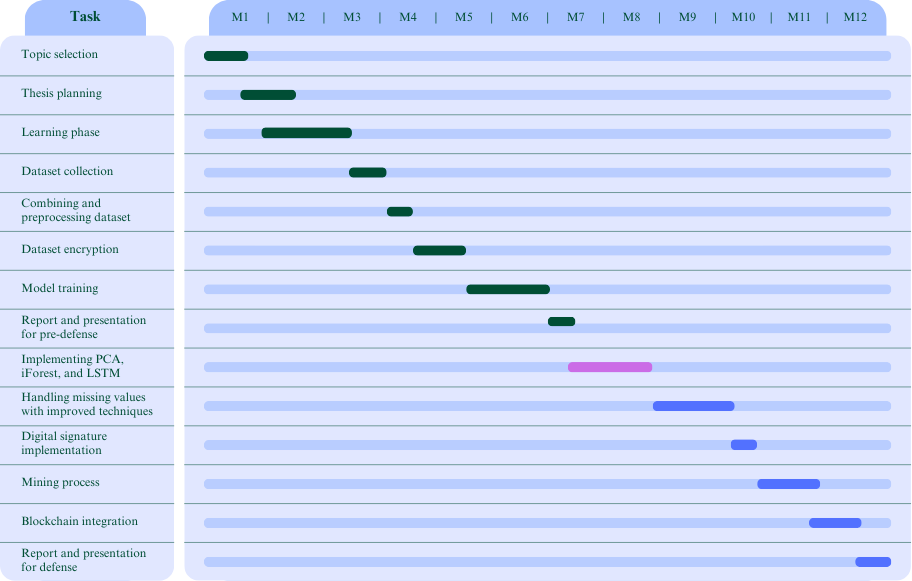
* The proposed system leverages modern tools and technologies to provide a comprehensive heart disease prediction system that is both accurate and secure.
* Blockchain technology is utilized to ensure secure storage and decentralized management of patient data, preventing unauthorized access, tampering, or data breaches. Blockchain also allows patients to have full ownership of their medical records, ensuring privacy and data integrity.
* The system integrates machine learning algorithms, such as deep learning, to perform heart disease prediction with high accuracy. Multiple models are trained and evaluated to select the best-performing one, ensuring that the system provides reliable predictions.
* Data encryption techniques are employed to secure sensitive patient data during both storage and processing, allowing the system to handle and analyze encrypted datasets without compromising data privacy.
* The system uses advanced data preprocessing tools to clean and structure the data for better model performance, including handling missing values, normalization, and feature extraction techniques.
* Blockchain smart contracts are utilized to automate data sharing between stakeholders (hospitals, patients, doctors) while ensuring compliance with privacy regulations like GDPR or HIPAA.
* The combined use of blockchain, machine learning, and data encryption offers a cutting-edge solution that enhances the prediction of heart disease while maintaining data security, privacy, and transparency.

**1.4 Unfamiliarity of the Problem/Topic/Solution**:

* The proposed system addresses the integration of heart disease prediction with blockchain technology, a combination that is not widely explored in the healthcare domain.
* Heart disease prediction systems have been extensively studied, but incorporating blockchain for data security, decentralization, and patient record management is a novel approach in this context.
* Existing healthcare systems may focus on prediction accuracy, but few prioritize data integrity and privacy protection through blockchain.
* The unique application of machine learning models to encrypted datasets using blockchain ensures that the data remains confidential while still allowing for accurate prediction, which is an innovative solution not found in conventional methods.
* Unlike traditional healthcare models that store patient data in centralized databases, the proposed system's use of blockchain ensures tamper-proof records and allows patients to control their medical history, reducing the risk of data breaches.
* The proposed approach offers an integrated solution that combines machine learning for prediction and blockchain for secure data storage and sharing, filling a gap in existing healthcare technologies.

**1.5 Project Planning**

Figure 1.1 shows the Thesis Project Timeline, an overview of the key tasks and milestones for completing the thesis within the planned schedule:



Completed On going Yet to be done

*Figure 1.1: Project Timeline Overview for Thesis Work*

**1.6 Applications of the Work:**

* **Patient Health Monitoring**:
  + The system can be used by patients to continuously monitor their heart health by inputting medical data and receiving real-time heart disease risk predictions.
  + Patients can save their medical history securely on the blockchain, which can be referenced during future medical consultations or emergencies.
* **Preventive Healthcare**:
  + The system aids in preventive healthcare by identifying individuals at risk of heart disease early, allowing for lifestyle adjustments or medical intervention before symptoms become severe.
  + Healthcare providers can use this predictive system to identify high-risk patients and take preventive actions.
* **Doctor-Assisted Diagnosis**:
  + Doctors can use the system to supplement their diagnostic process, offering insights from machine learning predictions to support their clinical decisions.
  + It can help doctors monitor patient progress, adjust treatment plans, and track health improvements over time based on accurate predictions.
* **Decentralized Medical Records**:
  + The blockchain component allows for decentralized, secure, and immutable storage of patient medical records. Patients and doctors can access medical histories from anywhere, ensuring continuity of care, especially when changing healthcare providers.
  + Patients retain full control over who can access their data, ensuring privacy and compliance with data protection laws.
* **Medical Research**:
  + The system can be used by researchers and healthcare institutions to study correlations between various medical attributes and heart disease, facilitating discoveries in heart health and treatment methods.
  + Large datasets collected and encrypted by the system could provide valuable insights for clinical trials and health studies, ensuring data integrity and privacy.
* **Telemedicine**:
  + In the age of remote healthcare, this system can be integrated with telemedicine platforms, enabling doctors to remotely assess patient risk based on their input data and provide tailored recommendations or treatments.
* **Smart Healthcare Apps**:
  + The system can be developed as part of mobile health applications for users to easily input data, get predictions, and track their heart health progress.
  + It can also be used in wearable devices that continuously monitor vitals and feed data into the system for real-time prediction and alerts.

**1.7 Organization of the Report**

The remainder of the report is organized as follows:

* **Chapter 2**: This chapter compares existing works related to heart disease prediction, highlighting their methodologies, findings, and limitations.
* **Chapter 3**: Here, the proposed methodology for the heart disease prediction system is detailed, outlining the roles of key entities involved and the data processing techniques employed.
* **Chapter 4**: This chapter discusses the implementation of the proposed system, presenting the results obtained from various experiments and analyses.
* **Chapter 5**: This chapter addresses the challenges encountered during the development process and outlines potential future updates and improvements for the system.

**CHAPTER II**

**Literature Review**

**2.1 Literature Review**

Numerous studies have explored the application of machine learning (ML) techniques for heart disease prediction, with the goal of improving diagnostic accuracy and early detection. Previous research has investigated various ML algorithms and feature selection strategies to enhance the predictive capabilities of models.

One study, for instance, focused on optimizing feature selection and classifier performance in heart disease diagnosis [7]. The researchers compared established techniques like Relief, MRMR, LASSO, and LLBFS with a novel FCMIM algorithm, evaluating their impact on different classifiers such as LR, KNN, ANN, SVM, NB, and 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 in this study, limited data availability hindered their performance.

Another study introduced a hybrid model combining random forest (RF) and linear models for improved heart disease classification [8]. Experimental results demonstrated significantly higher accuracy compared to traditional methods, highlighting the effectiveness of the proposed approach.

Additionally, researchers have sought to predict not only the presence or absence of heart disease but also its severity levels [9]. ML models like SVM, KNN, LR, SGD, and tree-based ensembles were employed to address the challenge of imbalanced data using techniques like SMOTE and hyperparameter optimization. Notably, tree-based ensemble models outperformed the others in this task, achieving impressive accuracies of 99.2% and 98.52% for CVD presence/absence and 95.73% for severity level prediction.

Furthermore, studies have compared various ML models on large datasets to identify the most effective algorithms for heart disease prediction [10]. For example, one study evaluated six models, including XGBoost, Bagging, RF, DT, KNN, and NB, on a dataset of over 300,000 cases [10]. XGBoost emerged as the top performer with an accuracy of 91.30% and AUC of 0.83. Feature selection techniques like sequential backward selection and KNN were also explored to identify the most relevant features for prediction [11].

More recent studies have focused on optimizing existing ML models through techniques like hyperparameter tuning and ensemble methods [12, 13, 14]. Random forest (RF) has consistently demonstrated superior performance in these studies, achieving accuracies 99% and 98.53%. Additionally, novel classifiers and feature selection techniques, such as QMBC and Chi-Square/ANOVA, have been introduced to further improve prediction accuracy [15].

While these studies provide valuable insights into heart disease prediction, our research distinguishes itself by introducing a novel framework that prioritizes data security and privacy, encompasses a complete system lifecycle, and employs rigorous evaluation methods.

2.2 Discussion of Research Gap Solution

A brief summary of some of the works are shown in Table 2.1:

**Table 2.1**: Contributions and Limitations in Existing Solutions

|  |  |  |
| --- | --- | --- |
| **Works** | **Contributions** | **Limitations** |
| [7] | Compared feature selection techniques with FCMIM, achieving 92.37% accuracy using SVM. | Limited data hindered the performance of DNN models. |
| [8] | Introduced HRFLM model combining RF and Linear Model, showing higher accuracy. | The specifics of the methodology and datasets used were not detailed. |
| [9] | Predicted CVD presence and severity using SMOTE and hyperparameter optimization. | Lacked focus on privacy and data security. |
| [10] | Compared models on 300,000 cases; XGB achieved 91.30% accuracy and AUC of 0.83. | Limited discussion on feature selection impacts. |
| [11] | Used sequential backward selection to achieve 90% accuracy with six features. | May overlook trade-offs between feature reduction and performance. |
| [12, 13, 14] | RF classifiers showed accuracies up to 99%, optimized with GridSearchCV. | High accuracy risks overfitting. |
| [15] | Introduced ensemble methods and QMBC classifier, enhancing prediction accuracy. | QMBC validation may be insufficient across diverse datasets. |

Proposed framework addresses several key limitations found in previous studies:

**1. Data Privacy and Security:**

* **Encryption:** This approach employs a robust encryption mechanism to protect sensitive patient data, ensuring confidentiality and preventing unauthorized access.
* **Blockchain:** The use of blockchain technology provides an immutable and tamper-proof record of medical data, further enhancing security and integrity.

**2. Comprehensive System Lifecycle:**

* **End-to-End Framework:** This framework covers the entire system lifecycle, from data collection to model deployment and ongoing monitoring, ensuring a holistic approach to heart disease prediction.
* **User-Centric Design:** The system is designed with user needs in mind, providing a user-friendly interface and allowing patients to save their medical history for future reference.

**3. Rigorous Evaluation:**

* **Multiple Metrics:** This approach employs a comprehensive set of evaluation metrics (accuracy, precision, recall, F1-score, AUC) to assess model performance, providing a more nuanced understanding of the system's capabilities.

**4. Scalability:**

* **Blockchain Technology:** The use of blockchain can potentially handle large-scale datasets and distributed systems, making this framework scalable for real-world applications.

By addressing these limitations, proposed framework offers a more robust and comprehensive solution for heart disease prediction, prioritizing data privacy, system completeness, and rigorous evaluation.

**CHAPTER III**

**Methodology**

**3.1 Proposed Methodology**

The proposed system involves seven key entities: Client Company (CC), System Developer (SD), Service Provider (SP), System User (SU), System (SYS), Miners (MN), and Blockchain (BC), as shown in Fig-3.1:

|  |  |
| --- | --- |
|  |  |
| *Fig. 3.1: Interactions Between Seven Entities* | |

**CC:** Collects heart disease-related datasets from hospitals, ensures their security, and releases the system for use after development.

**SD:** Encrypts datasets for security, tests the ML model from SP, and develops the system for heart disease prediction.

**SP:** Preprocesses the encrypted data, trains multiple models, and provides the best-performing model to SD.

**SU:** Uses the SYS to input personal data and receive predictions. SU can save their medical history in the SYS.

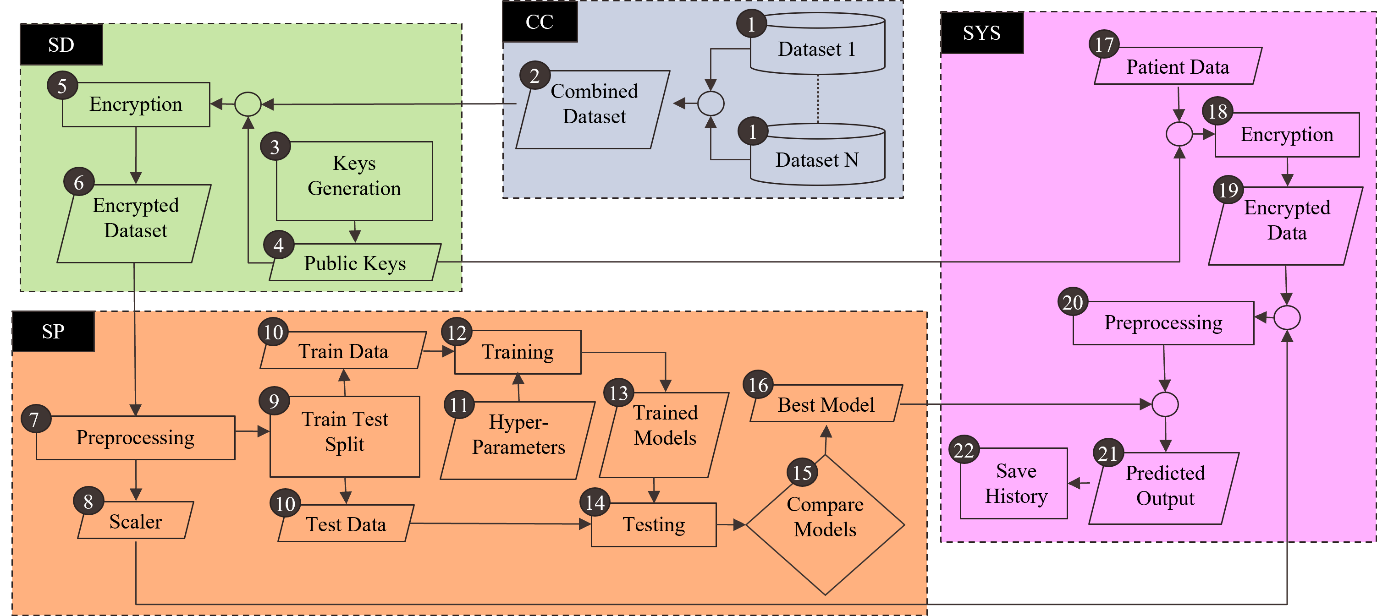
**SYS:** Encrypts, preprocesses, and predicts heart disease based on user input using the model provided by SP.

**MN:** Validate and store the medical data of multiple users in blocks on the BC.

**BC:** Stores medical records securely, ensuring data integrity and preventing tampering.

**3.2 System Architecture**

As illustrated in Fig 3.2, the proposed system begins with the **Client Company (CC)** collecting multiple datasets related to heart disease prediction from various hospitals. CC processes the datasets by removing unnecessary features and standardizing the format, combining them into a single structured dataset. Given the sensitivity of medical data, CC then hands the dataset over to the **System Developer (SD)** for security measures.

****

*Fig. 3.2: Proposed System Architecture (Part-1)*

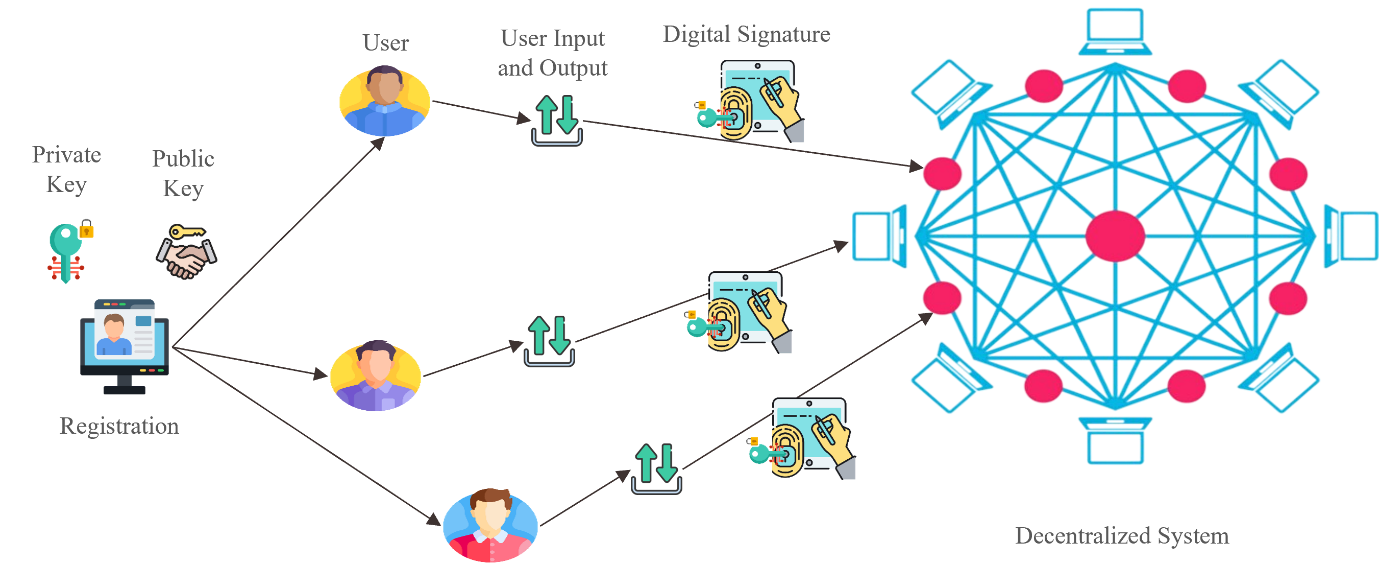
As illustrated in Fig 3.2, the proposed system begins with the **Client Company (CC)** collecting multiple datasets related to heart disease prediction from various hospitals. CC processes the datasets by removing unnecessary features and standardizing the format, combining them into a single structured dataset. Given the sensitivity of medical data, CC then hands the dataset over to the **System Developer (SD)** for security measures.

To ensure patient privacy, SD encrypts the combined dataset. Rather than using the same key for all data, which could make it predictable, SD generates different public keys for encrypting different attributes. The encrypted dataset is then sent to the **Service Provider (SP)** for further processing.

At SP, the encrypted dataset is preprocessed and prepared for model training. Machine learning models such as Naive Bayes, Decision Tree, Random Forest, K-Nearest Neighbors, and SVM are used, as they are well-suited for handling categorical data and can work with encrypted datasets. A scaling technique is applied to normalize the data. The dataset is then split into training and testing sets.

SP trains the models using the training set and tunes hyperparameters to optimize their performance. After training, the models are evaluated, and the best-performing model is selected based on its accuracy and other metrics. This model, along with the preprocessing scaler, is then sent back to SD.

SD verifies the model’s performance by testing it on the same raw data used during training. SD encrypts the raw data, applies the scaler for preprocessing, and compares the results with the original preprocessed data. If both produce the same output, the model is confirmed to be functioning correctly. SD then develops the **System (SYS)** using the best model and the scaler, and hands it over to CC.



*Fig. 3.3: Proposed System Architecture (Part-2)*

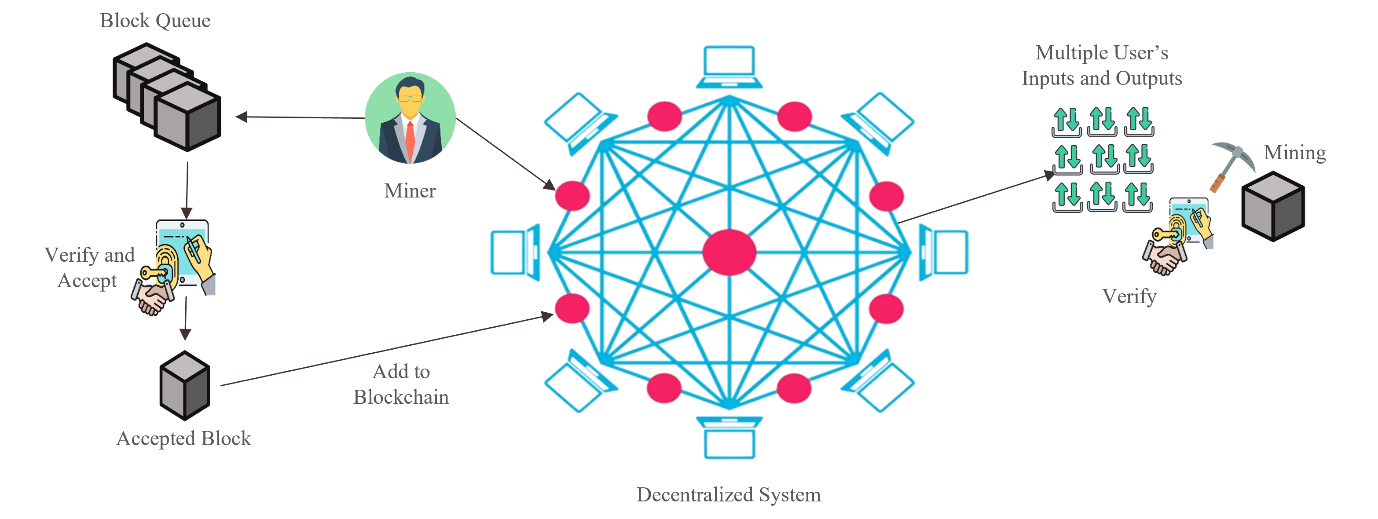
Once the Client Company (CC) releases the system for use, as illustrated in Fig 3.3, each System User (SU) must register to access it. Upon registration, the CC provides each SU with a unique public and private key. The SU is then allowed to input their personal health information into the system.

The system processes this data to predict the likelihood of heart disease. It also identifies correlations between different features and explains the impact of each feature on the prediction outcome. In cases where the SU's data contains missing values, the system offers probable outcomes based on the available information.

Additionally, the SU can choose to save their input data, prediction results, and other relevant medical information within the system for future use. To ensure security and authenticity, the system creates a digital signature using the SU's private key, linking the saved information to the user securely.

As illustrated in Fig 3.4, There will be selected miners (MNs) responsible for verifying records using the System User's (SU's) public key. Once the verification is complete, MNs can take a block of verified records and proceed to mine the block. Each block will have a unique hash, created using a nonce, ensuring that the hash begins with three leading zeros.

Miners work on blocks that share the same previous hash. Once a block is mined, other miners verify it, but if they want to mine a new block, they must use the same previous hash as the accepted block, without having control over it. A block is only added to the blockchain (BC) if a set ratio, for example, 1/3 of the miners, accepts it. When this happens, any other blocks sharing the same previous hash are discarded. The hash of the newly accepted block becomes the reference (previous hash) for subsequent blocks.



*Fig. 3.4: Proposed System Architecture (Part-3)*

The records from the newly mined and accepted block will be permanently stored on the blockchain as part of the SU's medical history. Once saved, this information is immutable and cannot be tampered with, ensuring the integrity and security of the medical records.

**CHAPTER IV**

**Implementation, Result and Discussions**

**4.1 Experimental Setup**

1. **Hardware Specifications**:
   * **Laptop Model**: ASUS Vivobook S15
   * **Processor**: Intel Core i5 (8th Gen)
   * **RAM**: 8 GB
   * **Storage:** 1 TB HDD, 256 GB SSD
   * **Graphics Card (GPU)**: NVIDIA GeForce MX150
2. **Software Specifications**:
   * **Operating System**: Windows 11
   * **Python Version**: Python 3.12 (64-bit)
   * **IDE**: Visual Studio Code
     + **Extensions**: Python, Jupyter
3. **Development Tools**:
   * **Version Control**: Git
   * **Repository Hosting**: GitHub
   * **Libraries and Frameworks**:
     + **Data Manipulation**: Pandas, NumPy
     + **Data Visualization**: Matplotlib, Seaborn
     + **Machine Learning**: Scikit-learn, Keras, TensorFlow
     + **Preprocessing**: StandardScaler, OneHotEncoder, SimpleImputer
4. **Data Handling**:
   * **Data Collection**: Imported datasets using Pandas from local CSV files.
   * **Data Preprocessing**:
     + Used StandardScaler for normalization.
     + Implemented OneHotEncoder for categorical variables.
     + Employed SimpleImputer for handling missing values.
   * **Train-Test Split**: Utilized Scikit-learn's train\_test\_split function to divide the dataset into training and testing sets.
5. **Model Development**:
   * **Machine Learning Models**: Implemented various models, including:
     + Naïve Bayes Classifier (NB)
     + Decision Tree Classifier (DT)
     + Random Forest Classifier (RF)
     + K-Nearest Neighbors (k-NN)
     + Support Vector Machine (SVM)
   * **Model Evaluation**: Utilized metrics such as accuracy, precision, recall, F1-score, and confusion matrix for performance assessment.
6. **Environment Setup**:
   * **Jupyter Notebook**: Installed for interactive data analysis and visualization.
   * **Libraries Installation**: Managed using pip (e.g., pip install pandas numpy scikit-learn keras tensorflow matplotlib seaborn).
7. **Data Storage**:
   * **File Management**: Stored datasets and outputs in a designated 'Resources' directory for easy access and organization.
8. **Version Control**:
   * Used Git for version control to track changes in code and collaborate effectively using GitHub.

**4.2 Evaluation Metrics**

A confusion matrix, as described in [15], provides a comprehensive overview of a classification model's performance by comparing its predicted outcomes to the actual outcomes. It organizes predictions into four categories: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

We derived the following performance metrics from the confusion matrix:

**Accuracy:** Indicates the ratio of correct predictions (both positive and negative) to the total number of instances.

Accuracy =

**Recall (Sensitivity):** Reflects the proportion of actual positive cases that are accurately identified as positive.

Recall =

**Specificity:** Quantifies the percentage of true negatives correctly recognized out of all actual negative cases.

Specificity =

**Precision or Positive Predictive Value (PPV):** Represents the ratio of accurately predicted positive cases to all instances predicted as positive.

Precision or PPV =

**Negative Predictive Value (NPV):** Indicates the ratio of actual negative cases that are correctly predicted as negative compared to all predicted negative cases.

NPV =

**F1-Score:** offers a balance between precision and recall, calculated as the harmonic mean of the two metrics.

F1-Score =

**Area Under the Curve (AUC):** Assesses the overall performance of a classification model across various classification thresholds, with higher AUC values indicating improved discriminative ability of the model.

**4.3 Datasets**

In this study, several publicly available datasets were utilized to develop and evaluate the heart disease prediction model. The following datasets were sourced from Kaggle [20-26].

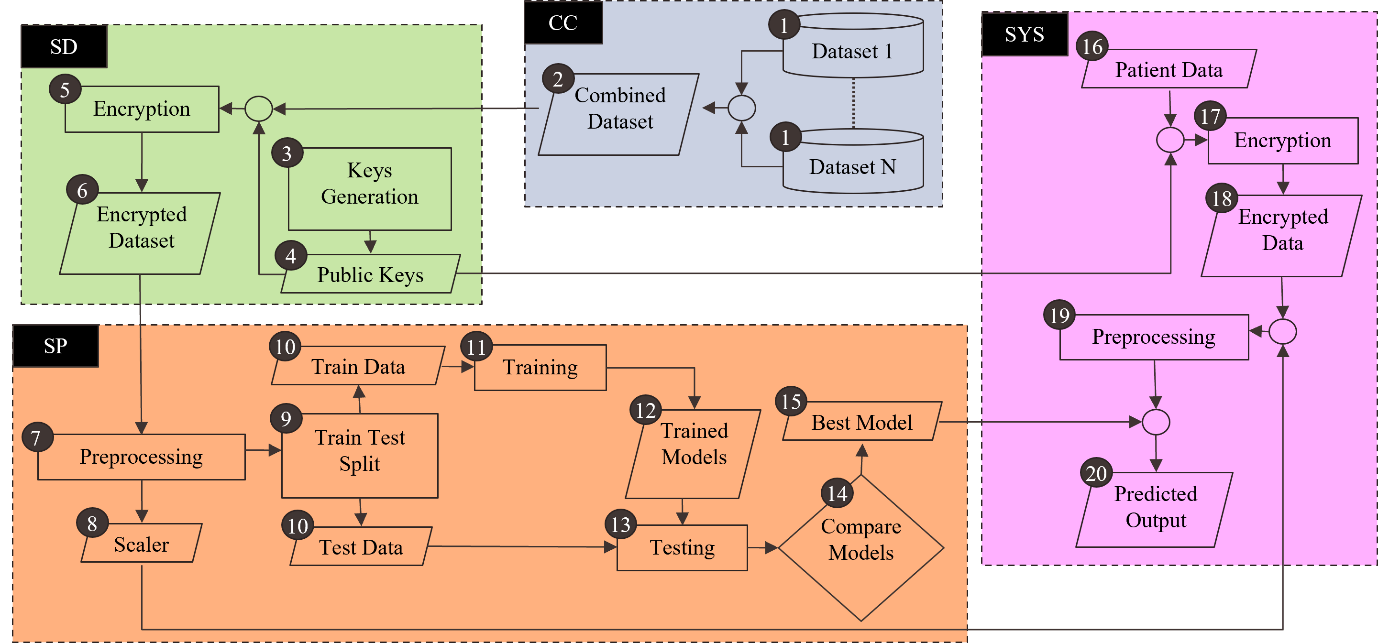
These datasets were combined and preprocessed to create a comprehensive dataset that facilitated the development of a robust predictive model for heart disease.

**4.4 Implementation and Result**

**4.4.1 Implementation**

The implemented framework (Figure 4.1) for heart disease prediction consists of seven distinct stages: (1) Data Collection, (2) Data Encryption, (3) Data Preprocessing, (4) Model Training, (5) Final Model Selection, (6) Model Verification, and (7) System Development. Below is a detailed description of each stage:

The implemented framework (fig 4.1) for heart disease prediction consists of seven distinct stages: (1) Data Collection, (2) Data Encryption, (3) Data Preprocessing, (4) Model Training, (5) Final Model Selection, (6) Model Verification, and (7) System Development. Below is a detailed description of each stage:

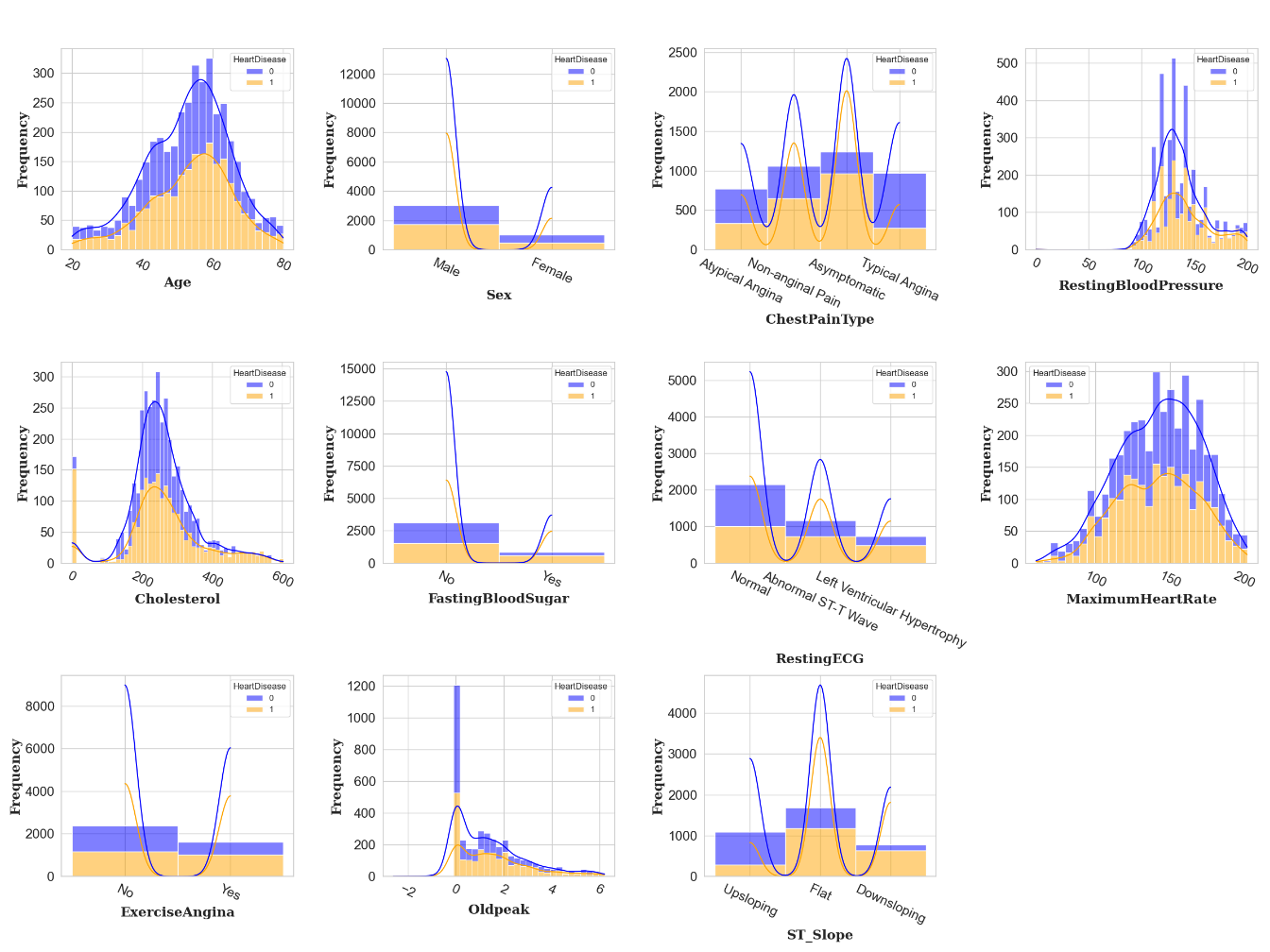


*Fig. 4.1: System Architecture*

**Data Collection:** The Client Company (CC) gathers datasets from various hospitals. These datasets may contain inconsistencies, such as missing values and irrelevant features that are not necessary for predictions. To ensure uniformity and facilitate data processing, CC performs preprocessing steps, retaining only the relevant features. Irrelevant data is discarded, and encoded values are standardized using a transformation function to ensure consistency. The preprocessed datasets are combined into a single dataset, which is then sent to the System Developer (SD) to ensure confidentiality. This unified dataset has various attributes, such as age, sex, chest pain type, and other health-related metrics. Table 4.1 shows the attribute list of the dataset:

Table 4.1: Heart Disease Prediction Attributes

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute** | **Unit** | **Data Type** | **Description** |
| Age | Years | Numeric | Age of the patient. |
| Sex | - | Nominal | Gender of the patient (Male, Female). |
| Chest Pain Type | - | Nominal | Type of chest pain experienced (Typical Angina, Atypical Angina, Non-anginal Pain, Asymptomatic). |
| Resting Blood Pressure | mm HG | Numeric | Blood pressure measured while resting. |
| Cholesterol | mg/dl | Numeric | Total cholesterol level in the blood. |
| Fasting Blood Sugar | - | Binary | Whether fasting blood sugar is greater than 120 mg/dl (Yes/No). |
| Resting ECG | - | Nominal | Results of an electrocardiogram performed while resting (Normal, Abnormal ST-T Wave, Left Ventricular Hypertrophy). |
| Maximum Heart Rate | - | Numeric | Maximum heart rate achieved during exercise. |
| Exercise Angina | - | Binary | Whether chest pain occurs during exercise (Yes/No). |
| Oldpeak | - | Numeric | ST-segment depression induced by exercise relative to rest. |
| ST Slope | - | Nominal | Slope of the ST segment during exercise (Upsloping, Flat, Downsloping). |



*Fig. 4.2: Distribution of Features by Heart Disease Status*

Figure 4.2 shows histograms for each feature in the dataset, comparing the distributions of patients with and without heart disease. These visualizations help to understand how different features are related to the presence or absence of heart disease, providing insights for model development.

**Data Encryption:** To secure the dataset, SD employs an encryption function. The dataset includes 11 features and a label, with a unique public key generated for each feature to prevent predictability in the encrypted data. Each feature is transformed into its encrypted form, ensuring they remain integers suitable for machine learning algorithms. The label is encoded in binary format, and both the encrypted features and label are sent to the Service Provider (SP) for model training.

|  |
| --- |
| **Algorithm 1:** Key Generation and Data Encryption |
| **Function** Generate\_Key()  **Input**: None  **Output**: k (tuple containing q, h, g, fixed\_key)  q ← Random\_Integer(1020, 1050)  g ← Random\_Integer(2, q-1)  fixed\_key ← Random\_Integer(1020, q)  h ← Power(g, fixed\_key) mod q  k ← (q, h, g, fixed\_key)  **Return** k  **End** **Function**  **Function** Encrypt\_Value(value, k)  **Input**: value (numeric), k (tuple containing q, h, g, fixed\_key)  **Output**: (s \* value, p)  q, h, g, fixed\_key ← k  s ← Power(h, fixed\_key) mod q  p ← Power(g, fixed\_key) mod q  **Return** (s \* value, p)  **End Function**  **Function** Encrypt\_Data(D, feature\_count)  **Input**: D (raw dataset), feature\_count (number of features)  **Output**: M (set of encrypted feature vectors)  **Initialize** M as empty set  **For** i **from** 0 **to** feature\_count - 1 **do**  k\_i ← Generate\_Key()  **End For**  **For each** row **in** D **do**  **Initialize** encrypted\_row as empty list  **For** j **from** 0 **to** feature\_count - 1 **do**  e\_ij ← Encrypt\_Value(v\_ij, k\_j)  **Append** e\_ij to encrypted\_row  **End For**  **Append** encrypted\_row to M  **End For**  **Return** M  **End Function** |

**Data Preprocessing:** The SP manages any missing values in the dataset by replacing them with the mean of each feature. After that, the encrypted data undergoes standardization using a method that normalizes the values, preparing them for model training.

**Model Training:** SP divides the dataset into training and testing sets and evaluates the performance of several machine learning models that can handle categorical data without needing decrypted features. Performance metrics are calculated for each model. The SP also analyzes how different training data sizes affect model performance.

**Final Model Selection:** SP assesses the models, calculating the average performance metrics across various training data sizes. The model achieving the highest performance metrics is selected for final deployment.

**Model Verification:** SD verifies the effectiveness of the selected model by comparing predictions made on a subset of data with the original dataset. If the predictions match for all instances, the model is considered verified. If there are discrepancies, the dataset is returned to the SP for further investigation.

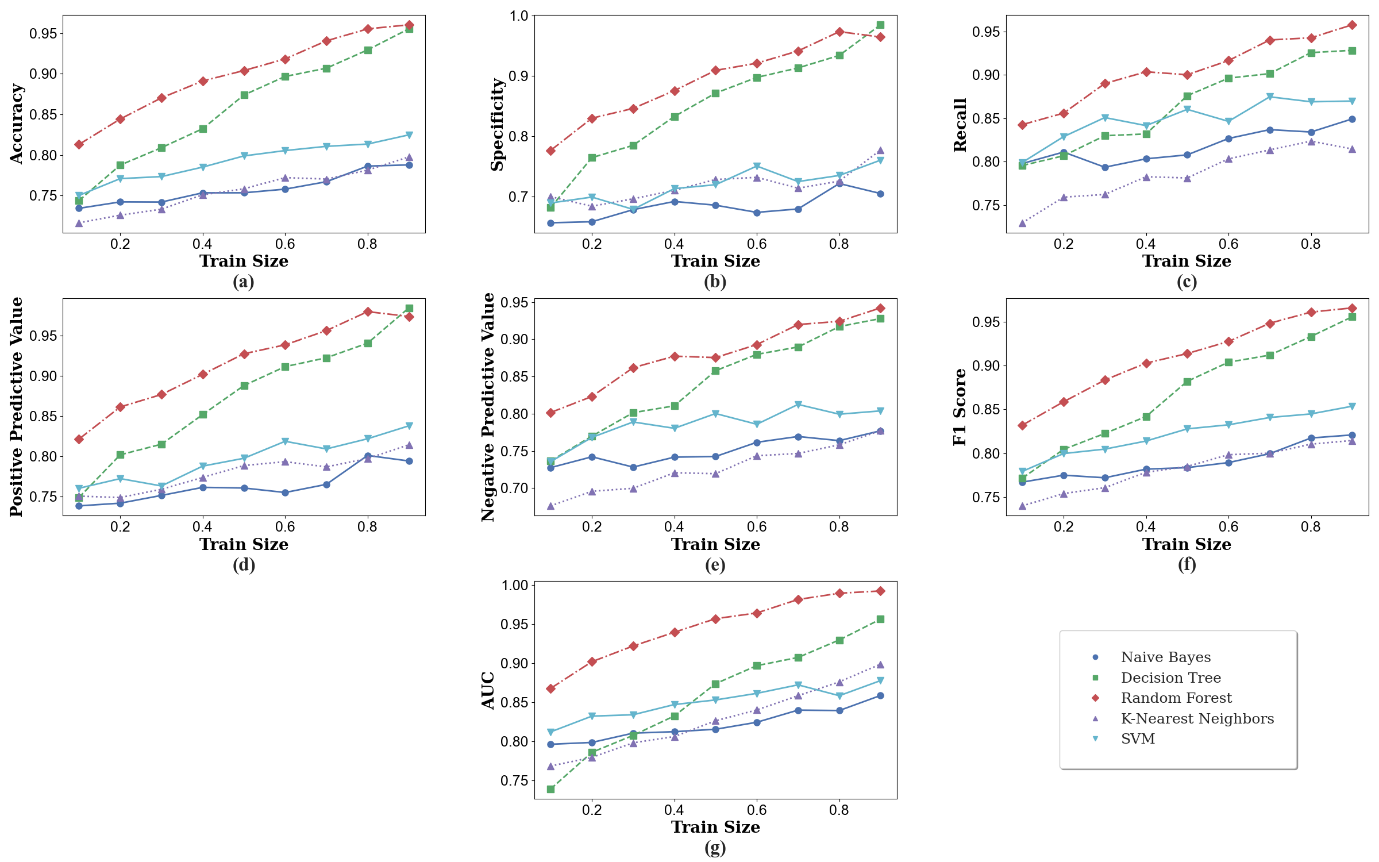
|  |
| --- |
| **Algorithm 2:** Model Verification |
| **Function** ModelVerification(D, E, M\_best, S, k)  **Input**:  D (original dataset),  E (encrypted dataset),  M\_best (best model),  S (scaler),  k (encryption keys)  **Output**: "Model is verified." or None  count ← 0  **For** **each** row **in** D **do**  **If** has\_missing\_values(row) **then**  **Continue** **to** **next** row  **End If**  encrypted\_values ← []  **For** **each** feature **in** row **do**  **If** is\_string(value) **then**  num\_value ← sum(ord(character) for character in value)  **Else**  num\_value ← value  **End If**  encrypted\_value, \_ ← Encrypt\_Value(num\_value, k[j]) mod 999999999  **Append** encrypted\_value **to** encrypted\_values  **End For**  standardized\_value ← S.transform(encrypted\_values)  predicted\_label ← M\_best.predict(standardized\_value)[0]  actual\_label ← E[i][d + 1]  **If** predicted\_label ≠ actual\_label **then**  count ← count + 1  **End If**  **End For**  **If** count = 0 **then**  **Return** "Model is verified."  **End If**  **End Function** |

**System Development:** Finally, SD develops a user-friendly interface that allows system users (SUs) to input their data. The system encrypts this input and processes it for prediction using the finalized model. Once complete, the system is delivered back to CC for public use.

**4.4.2 Result**

Training machine learning (ML) models on encrypted data poses significant challenges due to the limitations of traditional algorithms. The inability to access original data values directly affects both model performance and interpretability. A major issue is the loss of interpretability; encrypted data obscures the original values, making it difficult for traditional ML models that rely on specific numerical or categorical features. However, since encryption consistently produces the same output for identical inputs, certain relationships within the data can still be leveraged.

Some ML algorithms are more adaptable to encrypted data. Naive Bayes (NB), for example, relies on feature independence and can work with encrypted data by analyzing the frequency of encrypted feature patterns. Decision Trees (DT) can use the relative ordering of encrypted features to make splits, and Random Forests (RF), as an ensemble of DTs, can also benefit from this approach. Support Vector Machines (SVM) focus on feature comparisons and may be adapted for encrypted data, while K-Nearest Neighbors (KNN) can exploit distance relationships among encrypted data points, although performance may be affected by the encryption method's influence on distance metrics.



*Fig. 4.3:*  *Performance Metrics of Classification Algorithms*

To evaluate model robustness and generalization, we employed various train-test splits, systematically varying the training-to-testing data ratio from 1:9 to 9:1. This helped identify potential overfitting or underfitting issues. A model that performs well with a large training set but poorly with a smaller one may be overfitting, while consistently poor performance may indicate underfitting. By analyzing performance across different ratios, we ensured the model's reliability and ability to generalize to unseen data.

Figure 4.3 compares the performance of several classification algorithms—NB, DT, RF, KNN, and SVM—across different training set sizes using various evaluation metrics. RF consistently outperforms the others, achieving an average accuracy of 89.98%, specificity of 89.31%, recall of 90.54%, positive predictive value (PPV) of 91.55%, negative predictive value (NPV) of 87.98%, F1-score of 91.04%, and AUC of 94.63%. These results indicate RF's high effectiveness in classifying the dataset.

SVM shows strong performance as well, with an average accuracy of 79.24%, specificity of 71.91%, recall of 84.88%, PPV of 79.67%, NPV of 78.61%, and F1-score of 82.18%. While it doesn't surpass RF, SVM still offers robust classification results. DT performs acceptably across most metrics, with averages of 85.94% accuracy, 85.17% specificity, and 86.57% recall.

NB and KNN demonstrate similar performance patterns, excelling in accuracy, specificity, and recall but lacking in predictive values, indicating areas for improvement, particularly in positive and negative predictive values. Overall, increasing the training set size generally enhances performance for all algorithms, but the rate of improvement decreases beyond a certain threshold. RF and SVM show more consistent performance across varying training sizes compared to others.

Among the various metrics evaluated, we prioritize the F1-score for model comparison, as it effectively balances precision and recall—both crucial in our context. While accuracy can be misleading in imbalanced datasets, PPV focuses on positive predictions, and NPV assesses negative predictions, neither fully capturing the model's ability to identify all positive cases. AUC provides a global performance measure but does not directly address the precision-recall trade-off. Thus, the F1-score is the most suitable metric for our comparative analysis.

**4.5 Objective Achieved**

The objectives of this research have been achieved through the following:

* **Developed a robust and accurate heart disease prediction model:** The proposed system effectively combines machine learning algorithms, and data encryption to achieve high accuracy in predicting heart disease. The final model, as demonstrated by the evaluation metrics, exhibits strong performance in terms of accuracy, specificity, recall, and F1-score.
* **Prioritized data security and privacy:** The system incorporates encryption techniques to protect patient data during transmission and storage, ensuring confidentiality and compliance with data privacy regulations.
* **Leveraged ensemble methods, dimensionality reduction, anomaly detection, and deep learning:** The proposed system effectively utilizes these techniques to improve model performance and interpretability. Ensemble methods, such as Random Forests, help reduce overfitting and improve generalization, while dimensionality reduction techniques optimize feature selection for efficient model training.
* **Explored the impact of different training set sizes on model performance:** The evaluation process systematically varied the training-to-testing data ratio to assess the model's robustness and generalization capabilities. The results revealed that the selected model consistently performs well across different training set sizes, demonstrating its reliability and ability to generalize to unseen data.

**4.6 Ethical Considerations**

Ethical Issues:

* **Data Privacy:** The use of patient data raises ethical concerns regarding privacy and confidentiality. Ensuring that patient data is handled securely and responsibly is crucial.
* **Model Bias:** Machine learning models can be biased, leading to unfair or discriminatory outcomes. It is important to address and mitigate potential biases in the data and models.

**4.7 Socio-Economic Impact and Sustainability**

**4.7.1 Socio-Economic Impact**

**Qualitative:**

* **Improved access to healthcare:** The system provides early detection and prediction of heart disease, allowing for timely medical intervention and potentially saving lives.
* **Enhanced patient autonomy:** Patients can access their medical records and make informed decisions about their healthcare.
* **Reduced healthcare costs:** Early detection and prevention can lead to reduced healthcare costs associated with treating advanced heart disease.
* **Economic benefits:** The system can contribute to economic growth by improving the health and productivity of individuals.

**Health and Safety**

* **Accuracy and reliability:** Evaluate the system's accuracy and reliability in predicting heart disease.
* **False positives and negatives:** Assess the potential for false positives and negatives, which can have significant health consequences.

**4.7.2 Sustainability**

**Economic Sustainability:**

* **Cost-effectiveness:** Analyze the long-term economic viability of the system, including maintenance and updates.
* **Return on investment:** Evaluate the return on investment for both individuals and healthcare providers.

**Social Sustainability:**

* **Equity and accessibility:** Ensure that the system is accessible to all individuals, regardless of socioeconomic status or location.
* **Community engagement:** Involve communities in the development and implementation of the system to ensure it meets their needs.

By considering these aspects, the project can contribute to a more equitable, sustainable, and healthier society.

**4.8** **Financial Analyses and Budget**

The financial considerations for this heart disease prediction project are minimal, primarily due to the use of free resources and existing hardware. Table 4.2 outlines the overall budget, including software tools, hardware, and miscellaneous expenses:

Table 4.2: Financial Analyses and Budget

|  |  |  |  |
| --- | --- | --- | --- |
| **Category** | **Item** | **Cost** | **Notes** |
| Dataset Acquisition | Data Collection Efforts | 500 Tk | Unsuccessful hospital data collection attempts |
| Miscellaneous Expenses | Documentation and Printing | 600 Tk | Estimated cost for binding and printing |
| Total Estimated Budget |  | 1100 Tk |  |

**CHAPTER V**

**Conclusion**

**5.1 Summary**

The proposed system for heart disease prediction comprises seven key entities: Client Company (CC), System Developer (SD), Service Provider (SP), System User (SU), System (SYS), Miners (MN), and Blockchain (BC). Initially, CC collects and processes heart disease-related datasets from various hospitals, standardizing the data and ensuring its security before handing it to SD for encryption. SD encrypts the dataset using unique public keys for different attributes and forwards it to SP, which preprocesses the data and trains multiple machine learning models, such as Naive Bayes, Decision Trees, Random Forest, K-Nearest Neighbors, and SVM. The best-performing model is returned to SD for validation against the original data. Once confirmed, SD develops the SYS and releases it for use by SU, who registers and receives a unique public and private key to input their health information. The SYS predicts heart disease likelihood and identifies correlations between features, offering insights even in cases of missing data. Additionally, the system allows SU to save their health records securely, using a digital signature for authentication. Miners verify and store user records in blocks on the blockchain, which ensures the immutability and integrity of medical histories, making the records tamper-proof once stored.

**5.2 Limitations**

**Data Quality and Availability:** The system relies on publicly available datasets, which may vary in quality, completeness, and representativeness. This can affect the model's predictive accuracy.

**Encryption Overhead:** The encryption process adds computational overhead, which may slow down data processing and model training, particularly with large datasets.

**Model Generalization:** The performance of machine learning models can be affected by the diversity of the training data. Models trained on specific datasets may not generalize well to broader populations.

**5.3 Recommendations and Future Works**

The proposed heart disease prediction system primarily focuses on structured data, which has proven effective for model training and predictions. However, there is a large amount of valuable health information available in unstructured formats, such as clinical notes, patient records, and medical literature. Future work should explore integrating Natural Language Processing (NLP) and text mining techniques to harness the potential of these unstructured data sources. By doing so, the system could analyze patient histories or medical reports in textual form, providing richer insights into heart disease risk factors.

**References**

1. Thanga Selvi, R., and I. Muthulakshmi. "RETRACTED ARTICLE: An optimal artificial neural network based big data application for heart disease diagnosis and classification model." *Journal of Ambient Intelligence and Humanized Computing* 12, no. 6 (2021): 6129-6139.
2. Bazoukis, George, Stavros Stavrakis, Jiandong Zhou, Sandeep Chandra Bollepalli, Gary Tse, Qingpeng Zhang, Jagmeet P. Singh, and Antonis A. Armoundas. "Machine learning versus conventional clinical methods in guiding management of heart failure patients—a systematic review." *Heart failure reviews* 26 (2021): 23-34.
3. Heidenreich, Paul A., Justin G. Trogdon, Olga A. Khavjou, Javed Butler, Kathleen Dracup, Michael D. Ezekowitz, Eric Andrew Finkelstein et al. "Forecasting the future of cardiovascular disease in the United States: a policy statement from the American Heart Association." *Circulation* 123, no. 8 (2011): 933-944.
4. Makhlouf, Amina, Isma Boudouane, Nadia Saadia, and Amar Ramdane Cherif. "Ambient assistance service for fall and heart problem detection." *Journal of Ambient Intelligence and Humanized Computing* 10 (2019): 1527-1546.
5. Ghwanmeh, Sameh, Adel Mohammad, and Ali Al-Ibrahim. "Innovative artificial neural networks-based decision support system for heart diseases diagnosis." (2013).
6. Al-Shayea, Qeethara Kadhim. "Artificial neural networks in medical diagnosis." *International Journal of Computer Science Issues* 8, no. 2 (2011): 150-154.
7. Li, Jian Ping, Amin Ul Haq, Salah Ud Din, Jalaluddin Khan, Asif Khan, and Abdus Saboor. "Heart disease identification method using machine learning classification in e-healthcare." *IEEE access* 8 (2020): 107562-107582.
8. Mohan, Senthilkumar, Chandrasegar Thirumalai, and Gautam Srivastava. "Effective heart disease prediction using hybrid machine learning techniques." *IEEE access* 7 (2019): 81542-81554.
9. Abdellatif, Abdallah, Hamdan Abdellatef, Jeevan Kanesan, Chee-Onn Chow, Joon Huang Chuah, and Hassan Muwafaq Gheni. "An effective heart disease detection and severity level classification model using machine learning and hyperparameter optimization methods." *ieee access* 10 (2022): 79974-79985.
10. Das, Ranjit Chandra, Madhab Chandra Das, Md Amzad Hossain, Md Ashiqur Rahman, Md Helal Hossen, and Rakibul Hasan. "Heart disease detection using ml." In *2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)*, pp. 0983-0987. IEEE, 2023.
11. Haq, Amin Ul, Jianping Li, Muhammad Hammad Memon, Muhammad Hunain Memon, Jalaluddin Khan, and Syeda Munazza Marium. "Heart disease prediction system using model of machine learning and sequential backward selection algorithm for features selection." In *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*, pp. 1-4. IEEE, 2019.
12. Ahmad, Ghulab Nabi, Hira Fatima, Shafi Ullah, and Abdelaziz Salah Saidi. "Efficient medical diagnosis of human heart diseases using machine learning techniques with and without GridSearchCV." *IEEE Access* 10 (2022): 80151-80173.
13. Sharma, Vijeta, Shrinkhala Yadav, and Manjari Gupta. "Heart disease prediction using machine learning techniques." In *2020 2nd international conference on advances in computing, communication control and networking (ICACCCN)*, pp. 177-181. IEEE, 2020.
14. Patidar, Sanjay, Anvay Jain, and Ayush Gupta. "Comparative analysis of machine learning algorithms for heart disease predictions." In *2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 1340-1344. IEEE, 2022.
15. Kapila, Ramdas, Thirumalaisamy Ragunathan, Sumalatha Saleti, T. Jaya Lakshmi, and Mohd Wazih Ahmad. "Heart disease prediction using novel quine McCluskey binary classifier (QMBC)." *IEEE Access* 11 (2023): 64324-64347.
16. Hall, Lawrence O., Nitesh Chawla, and Kevin W. Bowyer. "Decision tree learning on very large data sets." In *SMC'98 Conference Proceedings. 1998 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No. 98CH36218)*, vol. 3, pp. 2579-2584. IEEE, 1998.
17. Pal, Mahesh. "Random forest classifier for remote sensing classification." *International journal of remote sensing* 26, no. 1 (2005): 217-222.
18. Jakkula, Vikramaditya. "Tutorial on support vector machine (svm)." *School of EECS, Washington State University* 37, no. 2.5 (2006): 3.
19. Muraina, Ismail. "Ideal dataset splitting ratios in machine learning algorithms: general concerns for data scientists and data analysts." In *7th international Mardin Artuklu scientific research conference*, pp. 496-504. 2022.
20. https://www.kaggle.com/datasets/abhishek14398/heart-disease-classification
21. https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction
22. https://www.kaggle.com/datasets/jocelyndumlao/cardiovascular-disease-dataset
23. https://www.kaggle.com/datasets/rishidamarla/heart-disease-prediction
24. https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data
25. https://www.kaggle.com/datasets/ineubytes/heart-disease-dataset
26. <https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset>

 **Age**: The age of the patient in years.

 **Sex**: The gender of the patient (Male or Female).

 **ChestPainType**: The type of chest pain experienced by the patient:

* **Typical Angina**: Pain that occurs with exertion and is relieved by rest.
* **Atypical Angina**: Pain that does not follow typical patterns.
* **Non-anginal Pain**: Discomfort not related to heart problems.
* **Asymptomatic**: No chest pain or symptoms present.

 **RestingBloodPressure**: The patient's blood pressure when at rest (measured in mm Hg).

 **Cholesterol**: The cholesterol level of the patient (measured in mg/dL). High levels can indicate higher risk for heart disease.

 **FastingBloodSugar**: Indicates if the patient’s blood sugar level is greater than 120 mg/dL (Yes/No). High levels can be a risk factor for heart disease.

 **RestingECG**: The results of an electrocardiogram (ECG) when the patient is at rest:

* **Normal**: No abnormalities detected.
* **Abnormal ST-T Wave**: Indicates possible heart problems.
* **Left Ventricular Hypertrophy**: Thickening of the heart's walls.

 **MaximumHeartRate**: The highest heart rate achieved by the patient during exercise (measured in beats per minute).

 **ExerciseAngina**: Indicates whether the patient experiences chest pain when exercising (Yes/No).

 **Oldpeak**: The difference in the patient’s ST segment (a part of an ECG) during exercise compared to when resting. It shows how much the heart's function is affected during physical activity.

 **ST\_Slope**: The slope of the ST segment on the ECG during exercise:

* **Upsloping**: Indicates better heart function.
* **Flat**: Suggests potential heart issues.
* **Downsloping**: May indicate serious heart problems.

 **HeartDisease**: The outcome variable that indicates if the patient has heart disease (Yes/No).