*Indian Engineering Journal*
*ISSN: 2984-9047*
*Vol 4 Issue 1 (2026)*

# Enhancing the Predictive Performance of Heart Disease Detection on The Cardiovascular Disease Dataset Using Gated Recurrent Unit and Light Gradient Boosting Machine

**[1]M. Ranjani, [2]Dr. P.R.Tamilselvi**

[1]Research Scholar, Department of Computer Science, Periyar University, Salem, TamilNadu, India.

[2]Assistant Professor, Department of Computer Science, Government Arts and Science College, (Affiliated to Periyar University), Komarapalayam,TamilNadu, India

**Abstract**
Heart disease is one of the leading causes of mortality worldwide, making early and accurate diagnosis essential for improving patient survival rates and reducing healthcare costs. Traditional diagnostic techniques often face challenges in achieving high predictive accuracy due to the complexity and variability of cardiovascular data. To address these limitations, this research proposes an intelligent heart disease prediction framework using two advanced approaches: Gated Recurrent Unit (GRU) and Light Gradient Boosting Machine (LightGBM). The proposed models are implemented and evaluated on the Cardiovascular Disease Dataset to enhance predictive performance and classification efficiency. Initially, the dataset undergoes preprocessing techniques including data cleaning, normalization, and feature preparation to improve data quality and model performance. The GRU model is employed to capture complex hidden patterns and temporal relationships within the dataset, while the LightGBM model is utilized for efficient gradient boosting-based classification with reduced computational complexity. The performance of both models is evaluated using Accuracy, Precision, Recall (Sensitivity), Specificity, and F1-Score. Experimental results demonstrate that the proposed models significantly improve heart disease prediction performance compared to conventional machine learning techniques. Among the evaluated models, LightGBM achieves superior classification efficiency and faster training performance, whereas GRU effectively learns complex feature representations from medical data.

**Keywords:** Heart Disease Detection, Cardiovascular Disease Dataset,Gated Recurrent Unit (GRU) , Light Gradient Boosting Machine (LightGBM) , Deep Learning , Machine Learning , Healthcare Analytics , Predictive Modeling , Classification , Medical Data Analysis

## I. Introduction

Cardiovascular diseases (CVDs) are among the most serious health disorders affecting millions of people worldwide. These diseases primarily involve abnormalities in the heart and blood vessels, including coronary artery disease, heart failure, arrhythmia, and stroke. Rapid urbanization, unhealthy lifestyles, poor dietary habits, smoking, obesity, diabetes, hypertension, and physical inactivity have significantly increased the prevalence of cardiovascular diseases in recent years. Due to the complexity of cardiac conditions and their associated risk factors, early diagnosis and continuous monitoring have become essential in modern healthcare systems. According to global health reports, cardiovascular diseases remain one of the leading causes of mortality worldwide, accounting for a significant number of deaths each year [1]. The increasing burden of heart-related diseases affects not only patient health but also healthcare infrastructures and economic stability. In many developing and developed countries, delayed diagnosis and insufficient access to advanced medical facilities contribute to high mortality rates. Therefore, accurate and timely prediction of heart disease is crucial for reducing fatality rates and improving patient survival. Early heart disease detection plays a vital role in preventive healthcare and clinical decision-making. Traditional diagnostic procedures often require extensive medical examinations, laboratory testing, and expert interpretation, which may increase diagnosis time and healthcare costs. In recent years, machine learning and deep learning techniques have emerged as effective solutions for intelligent disease prediction and medical data analysis. These approaches can automatically learn hidden patterns from healthcare datasets and assist physicians in identifying high-risk patients at an early stage [2].

Among various intelligent techniques, deep learning models such as Gated Recurrent Unit (GRU) and advanced boosting algorithms such as Light Gradient Boosting Machine (LightGBM) have shown promising performance in medical prediction tasks. GRU is capable of learning complex feature relationships and sequential patterns efficiently, while LightGBM provides faster training and higher classification accuracy with reduced computational complexity. Therefore, integrating these advanced predictive models for cardiovascular disease detection can significantly enhance healthcare analytics and support accurate medical diagnosis.

**1.1 Problem Statement:** Traditional heart disease diagnosis methods primarily rely on clinical observations, manual interpretation of medical reports, and physician expertise. Although these methods are widely used in healthcare systems, they often face limitations in terms of prediction accuracy, time consumption, and scalability. In many cases, conventional diagnostic approaches may fail to identify hidden relationships among multiple clinical parameters, leading to delayed or inaccurate diagnosis of cardiovascular diseases. Existing machine learning models used for heart disease prediction also encounter several challenges, including overfitting, high computational cost, imbalanced datasets, and insufficient feature learning capability. Some traditional algorithms are unable to effectively capture nonlinear relationships and complex dependencies among medical attributes. As a result, the prediction performance may decrease when handling large-scale or high-dimensional cardiovascular datasets. Furthermore, many existing studies focus primarily on improving accuracy while neglecting other important performance measures such as precision, recall (sensitivity), specificity, and F1-score. In medical diagnosis systems, sensitivity and specificity are highly important because incorrect predictions can directly affect patient treatment and survival outcomes. Therefore, there is a strong need for intelligent predictive models that can achieve balanced and reliable classification performance across multiple evaluation metrics. To overcome these limitations, this research proposes the use of Gated Recurrent Unit (GRU) and Light Gradient Boosting Machine (LightGBM) for enhancing heart disease prediction on the Cardiovascular Disease Dataset. The proposed approach aims to improve predictive accuracy, efficient feature learning, and overall classification performance while reducing computational complexity and improving healthcare decision support systems.

**1.2 Research Motivation:** The increasing number of cardiovascular disease cases worldwide has created a strong need for intelligent healthcare prediction systems. Traditional diagnosis methods are often time-consuming and depend heavily on physician expertise, which may reduce prediction accuracy. Advanced machine learning and deep learning techniques can efficiently analyze medical data and improve disease prediction. GRU provides effective feature learning with lower computational complexity, while LightGBM offers faster training and high classification accuracy. Therefore, this research focuses on implementing GRU and LightGBM models to improve heart disease prediction performance and support intelligent healthcare systems.

**1.3 Research Objectives:** The main objective of this research is to improve heart disease prediction using GRU and LightGBM models on the Cardiovascular Disease Dataset. The study aims to preprocess and analyze the dataset, implement GRU for deep feature learning, develop LightGBM for efficient classification, evaluate model performance using Accuracy, Precision, Recall, Specificity, and F1-Score, compare both models, and identify the best-performing approach for intelligent healthcare prediction.

**1.4 Research Contributions:** This research contributes to intelligent cardiovascular disease prediction by developing GRU and LightGBM-based prediction models. The study improves dataset quality through preprocessing techniques and performs comparative analysis using multiple evaluation metrics. The proposed framework enhances prediction accuracy, computational efficiency, and healthcare decision support for early heart disease diagnosis and clinical risk assessment.

The remainder of this paper is organized into several sections to present the proposed research systematically. **Section 1** presents the introduction, including the background of cardiovascular diseases, problem statement, research motivation, objectives, contributions, and overall organization of the paper. **Section 2** discusses the literature review related to heart disease prediction methods, machine learning techniques, deep learning approaches, and boosting algorithms used in previous research studies. **Section 3** describes the proposed methodology, including dataset description, preprocessing techniques, feature selection methods, implementation of the GRU and LightGBM models, and performance evaluation metrics. **Section 4** presents the experimental results and discussion, including training and testing analysis, comparative performance evaluation, graphical

analysis, and interpretation of obtained results. **Section 5** concludes the research by summarizing the major findings, highlighting the significance of the proposed work, and suggesting future enhancements for intelligent heart disease prediction systems.

## Ii. Literature Review

### 2.1 Existing Heart Disease Prediction Techniques

Heart disease prediction has become one of the major research areas in healthcare analytics due to the increasing prevalence of cardiovascular disorders worldwide. Researchers have applied several traditional machine learning algorithms to improve prediction accuracy and assist clinicians in early diagnosis. These techniques analyze medical datasets and identify hidden patterns associated with cardiovascular diseases.

Support Vector Machine (SVM) is one of the most commonly used machine learning algorithms for heart disease classification. SVM performs classification by constructing an optimal hyperplane that separates data into different classes with maximum margin. Machine Learning Khemphila et al., (2011) applied SVM for heart disease prediction and reported improved classification performance compared to traditional statistical methods. However, the algorithm faced challenges when handling large-scale and high-dimensional medical datasets due to increased computational complexity[3].

Random Forest is another widely used ensemble learning algorithm for medical prediction systems. It combines multiple decision trees to improve classification accuracy and reduce overfitting. Palaniappan et al., (2008) utilized Random Forest for cardiovascular disease prediction and demonstrated that the model achieved higher prediction reliability and better generalization performance than single classifiers. The algorithm also provided improved feature importance analysis for medical attributes [4].

Decision Tree algorithms are extensively employed in healthcare applications because of their simple structure and easy interpretability. Decision trees classify data using hierarchical decision rules derived from dataset features. Srinivas et al., (2010) proposed a Decision Tree-based heart disease prediction system and observed that the model produced understandable diagnostic rules for clinical interpretation. However, the model suffered from instability and overfitting issues when applied to complex datasets [5].

In recent years, XGBoost has gained considerable attention for medical prediction tasks due to its enhanced boosting capability and high computational efficiency. XGBoost utilizes gradient boosting techniques to improve classification accuracy while minimizing training errors. Chen et al., (2016) demonstrated that XGBoost achieved superior performance in predictive analytics because of its regularization mechanism and optimized tree learning process. Several healthcare studies have adopted XGBoost for disease prediction due to its robustness and scalability [6].

### 2.2 Deep Learning Approaches

Deep learning techniques have significantly improved healthcare analytics by enabling automatic feature extraction and efficient learning from complex medical datasets. Unlike traditional machine learning methods, deep learning models can learn hierarchical feature representations and capture hidden dependencies among clinical attributes.

Recurrent Neural Networks (RNNs) are widely used for sequential data analysis and predictive modeling. RNNs maintain memory states that allow the model to process temporal information and learn dependencies between input sequences. Lipton et al., (2015) applied RNN models for clinical prediction tasks and demonstrated their capability to analyze healthcare records effectively. However, traditional RNNs often suffer from vanishing gradient problems, which reduce learning efficiency for long-term dependencies [7].

To overcome the limitations of RNNs, Long Short-Term Memory (LSTM) networks were introduced with specialized memory cells and gating mechanisms. LSTM models effectively retain important information over long sequences and improve prediction accuracy. Rajkomar et al., (2018) utilized LSTM networks for large-scale

clinical data prediction and achieved significant improvements in medical diagnosis performance. The ability of LSTM to learn complex feature relationships makes it suitable for heart disease prediction applications [8].

Bidirectional Long Short-Term Memory (BiLSTM) models further enhance learning capability by processing data in both forward and backward directions. This dual processing mechanism improves contextual understanding and feature extraction. Siami-Namini et al., (2019) reported that BiLSTM models achieved higher prediction accuracy in sequential healthcare data analysis compared to conventional LSTM architectures [9].

Among recent deep learning architectures, Gated Recurrent Unit (GRU) has emerged as an efficient alternative to LSTM. GRU simplifies the memory structure by using update and reset gates while maintaining effective learning capability. Chung et al., (2014) demonstrated that GRU models achieved comparable performance to LSTM with lower computational complexity and faster training speed. In healthcare applications, GRU has shown excellent performance in disease prediction, patient monitoring, and medical sequence analysis [10].

Several studies have utilized GRU models for intelligent healthcare systems. Esteban et al., (2016) applied GRU-based architectures for clinical time-series prediction and observed improved sensitivity and prediction reliability. The efficient feature learning capability of GRU makes it highly suitable for cardiovascular disease detection using complex medical datasets [11].

**2.3 Boosting Algorithms in Medical Prediction**

Boosting algorithms have become highly effective in medical data classification because they combine multiple weak learners to produce strong predictive models. These methods improve classification accuracy and reduce prediction errors through iterative learning processes.

Gradient Boosting is a powerful ensemble learning technique that sequentially builds decision trees to minimize classification loss. Friedman et al., (2001) introduced the Gradient Boosting framework and demonstrated its effectiveness in predictive analytics and regression tasks. In healthcare applications, Gradient Boosting algorithms have been successfully applied to disease diagnosis and risk prediction due to their strong classification capability [12].

XGBoost is an advanced implementation of Gradient Boosting that incorporates regularization techniques and optimized parallel processing. Chen et al., (2016) proposed XGBoost as a scalable and efficient boosting algorithm capable of handling large-scale datasets with improved computational speed. In medical prediction systems, XGBoost has achieved superior performance in heart disease classification, diabetes prediction, and cancer diagnosis because of its ability to model complex nonlinear relationships [13].

Light Gradient Boosting Machine (LightGBM) is another advanced boosting framework developed to overcome computational limitations associated with traditional boosting methods. Ke et al., (2017) introduced LightGBM with a leaf-wise tree growth strategy and histogram-based learning mechanism, enabling faster training and lower memory consumption. The algorithm significantly improves prediction efficiency while maintaining high classification accuracy [14].

In healthcare applications, LightGBM has demonstrated excellent performance in medical image analysis, disease prediction, and patient risk assessment. Zhang et al., (2020) applied LightGBM for cardiovascular disease prediction and reported improved accuracy, sensitivity, and faster execution time compared to conventional machine learning approaches. Due to its scalability and efficient feature handling capability, LightGBM is considered highly suitable for intelligent healthcare analytics [15].

Another major limitation observed in existing studies is computational complexity. Traditional deep learning models such as RNN and LSTM require extensive computational resources and longer training times when handling large-scale healthcare datasets. Although advanced boosting algorithms improve efficiency, some models still struggle with feature selection and optimization in high-dimensional medical data environments. Furthermore, many existing heart disease prediction systems utilize single-model approaches without integrating advanced hybrid optimization techniques. Limited research has been conducted on combining efficient deep learning models with lightweight boosting frameworks to improve both predictive accuracy and computational performance

simultaneously. Existing studies also face challenges in handling nonlinear dependencies and extracting meaningful features from complex cardiovascular datasets. Therefore, there is a strong need for intelligent predictive frameworks that can improve sensitivity, specificity, computational efficiency, and feature learning capability for accurate heart disease detection. Motivated by these research gaps, the present study focuses on implementing GRU and LightGBM models for enhanced cardiovascular disease prediction.

Summary of Literature Review,

| **Author** | **Technique Used** | **Advantages** | **Limitations** |
|---|---|---|---|
| Khemphila et al., (2011) | SVM | Improved classification accuracy | High computational complexity |
| Palaniappan et al., (2008) | Random Forest | Better generalization performance | Increased model complexity |
| Srinivas et al., (2010) | Decision Tree | Easy interpretation | Overfitting issues |
| Chen et al., (2016) | XGBoost | High prediction efficiency | Parameter tuning complexity |
| Lipton et al., (2015) | RNN | Sequential data learning | Vanishing gradient problem |
| Rajkomar et al., (2018) | LSTM | Better long-term dependency learning | High training time |
| Siami-Namini et al., (2019) | BiLSTM | Improved contextual learning | Increased computational cost |
| Chung et al., (2014) | GRU | Faster training and lower complexity | Limited optimization studies |
| Ke et al., (2017) | LightGBM | Faster execution and low memory usage | Requires proper parameter tuning |
| Zhang et al., (2020) | LightGBM for healthcare | High accuracy and sensitivity | Limited hybrid implementation |

**Iii. Research Methodology**

The proposed methodology is designed to enhance the predictive performance of heart disease detection using two advanced intelligent models: Gated Recurrent Unit (GRU) and Light Gradient Boosting Machine (LightGBM). The overall framework consists of several sequential stages including dataset acquisition, preprocessing, feature selection, model training, prediction, and performance evaluation. The proposed system aims to improve classification accuracy and healthcare decision-making by efficiently analyzing cardiovascular medical data. Initially, the Cardiovascular Disease Dataset is collected and subjected to preprocessing operations to improve data quality and remove inconsistencies. Preprocessing steps include handling missing values, normalization, outlier removal, and feature encoding. After preprocessing, feature selection techniques are applied to identify the most significant medical attributes influencing heart disease prediction. The processed dataset is then divided into training and testing datasets[16][17]. The GRU model is employed to learn hidden feature relationships and nonlinear dependencies within the cardiovascular dataset, while the LightGBM model performs efficient gradient boosting-based classification with optimized computational performance. Finally, the prediction results are evaluated using Accuracy, Precision, Recall (Sensitivity), Specificity, and F1-Score.

### 3.1 Dataset Description

The proposed research utilizes the Cardiovascular Disease Dataset, which contains medical records related to heart disease diagnosis and patient health conditions. The dataset is widely used for healthcare analytics and cardiovascular disease prediction research. It consists of demographic, clinical, and physiological attributes that contribute to identifying the risk of heart disease [18]. The dataset includes approximately 70,000 patient records with multiple medical features collected from healthcare examinations. Each record represents the health condition of an individual patient, while the target class indicates the presence or absence of cardiovascular disease.

**Dataset Attributes**

| **Attribute** | **Description** |
|---|---|
| Age | Age of the patient |
| Gender | Male/Female |
| Height | Height of the patient |
| Weight | Weight of the patient |
| $Ap_{hi}$ | Systolic blood pressure |
| $Ap_{lo}$ | Diastolic blood pressure |
| Cholesterol | Cholesterol level |
| Glucose | Glucose level |
| Smoke | Smoking habit |
| Alcohol Intake | Alcohol consumption status |
| Physical Activity | Physical activity condition |
| Cardio | Target variable indicating heart disease |

The dataset contains both numerical and categorical attributes, making preprocessing and feature transformation essential before model implementation.

### 3.2 Data Preprocessing

Data preprocessing is an important step in machine learning and deep learning systems because raw medical datasets often contain missing values, noisy information, inconsistent scales, and redundant data. Proper preprocessing improves model performance, prediction accuracy, and computational efficiency.

**Handling Missing Values:** Missing values occur when certain attribute values are unavailable in the dataset. Missing data can negatively affect predictive performance and introduce bias into the model. Therefore, missing values are handled using mean imputation for numerical attributes.

$$X_{mean} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

Where, $X_{mean}$ represents the mean value of the attribute, $X_i$ denotes individual attribute values and N represents the total number of available observations.The missing values are replaced with the mean of the corresponding feature to maintain dataset consistency and reduce information loss [19].

**Data Normalization/Scaling:** Medical attributes often contain different value ranges, which may negatively affect model learning performance. Normalization transforms feature values into a common range to improve convergence and training stability.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Where, $X_{norm}$ is the normalized value, X represents the original feature value, $X_{min}$ is the minimum value of the feature and $X_{max}$ is the maximum value of the feature. This normalization method scales feature values between 0 and 1, improving model training efficiency and reducing dominance of larger numerical attributes [20].

**Outlier Removal:** Outliers are abnormal observations that significantly differ from normal data points. In healthcare datasets, outliers may occur due to measurement errors or inconsistent medical records. Removing outliers helps improve model reliability and prediction accuracy.

The Interquartile Range (IQR) method is used for outlier detection.

$$IQR = Q3 - Q1$$

Outlier Detection Boundary,

Lower Limit=Q1−1.5(IQR)

Upper Limit=Q3+1.5(IQR)

Where, Q1 represents the first quartile, Q3 represents the third quartile and IQR denotes the interquartile range. Any data point outside the lower and upper limits is considered an outlier and removed from the dataset.

**Feature Encoding:** Machine learning models require numerical input values; therefore, categorical features must be converted into numerical representations. Label Encoding is used to transform categorical variables into integer values.

Encoded Value=f(Category)

Where, f(Category) represents the mapping function that assigns numerical labels to categorical classes. For example:

| **Category** | **Encoded Value** |
|---|---|
| Male | 0 |
| Female | 1 |

Feature encoding enables efficient processing of categorical healthcare attributes within predictive models.

**3.3 Feature Selection**

Feature selection is used to identify the most relevant medical attributes influencing heart disease prediction. Selecting important features improves prediction accuracy, reduces computational complexity, and minimizes overfitting.

**Correlation Analysis:** Correlation analysis measures the relationship between input features and the target variable. Features with strong correlation values are considered significant for heart disease prediction.

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})\ \ \sum(Y_i - \bar{Y})^2}}$$

Where, r denotes the correlation coefficient , $X_i$ and $Y_i$ represent feature values and $\bar{X}$ and $\bar{Y}$ represent mean values.

The correlation coefficient ranges between -1 and +1:

- +1 indicates strong positive correlation
- −1 indicates strong negative correlation
- 0 indicates no correlation

Highly correlated medical features are selected for model training.

**Important Medical Attributes**

Based on correlation analysis and medical significance, important attributes influencing cardiovascular disease prediction include:

| **Attribute** | **Importance** |
|---|---|
| Age | Higher age increases heart disease risk |
| Blood Pressure | Indicates cardiovascular stress |
| Cholesterol | High cholesterol increases blockage risk |
| Glucose Level | Related to diabetes and heart conditions |
| Smoking Habit | Major cardiovascular risk factor |
| Physical Activity | Reduces heart disease risk |
| Weight and BMI | Obesity contributes to heart disease |

These selected features are used as input for GRU and LightGBM models to improve predictive performance and classification reliability.

### 3.4. GRU Model

The Gated Recurrent Unit (GRU) is a powerful deep learning architecture designed to process sequential and nonlinear data efficiently. GRU is a simplified variant of Long Short-Term Memory (LSTM) networks that uses gating mechanisms to control information flow and retain important features during training. In this research, the GRU model is implemented to learn complex medical patterns and improve heart disease prediction performance on the Cardiovascular Disease Dataset. The proposed GRU model consists of four major components are Input Layer, GRU Hidden Layers , Dense Layer and Output Layer. The architecture enables efficient feature learning, reduced computational complexity, and improved prediction accuracy for cardiovascular disease detection [22][23].

#### 3.4.1 GRU Architecture

**Input Layer:** The input layer receives preprocessed cardiovascular dataset features and forwards them to the GRU hidden layers for sequential learning. Each input sample contains medical attributes such as age, blood pressure, cholesterol, glucose level, smoking habit, and physical activity.

$$X = [x_1, x_2, x_3, \dots, x_n]$$

Where, X represents the input feature vector , $x_1, x_2, \dots, x_n$ denote medical attributes and n indicates the total number of features . The input layer transfers normalized feature vectors into the GRU network for further processing and pattern learning.

**GRU Hidden Layers:** The GRU hidden layers are responsible for extracting hidden feature relationships and learning temporal dependencies within the cardiovascular dataset. GRU utilizes gating mechanisms to selectively retain important information while discarding irrelevant data.

The hidden state of the GRU is computed using the following formula:

$$h_t = (1 - z_t) \quad h_{t-1} + z_t \quad \tilde{h}_t$$

Where, $h_t$ represents the current hidden state , $h_{t-1}$ denotes the previous hidden state , $z_t$ represents the update gate , $h_{\sim t}$ indicates the candidate hidden state and $\odot$ denotes element-wise multiplication . The GRU hidden layers effectively capture nonlinear dependencies among medical features and improve feature learning capability.

**Dense Layer:** The dense layer receives extracted features from the GRU hidden layers and performs nonlinear transformation for classification. It combines learned representations and prepares the final prediction output.

$$y = f(W_x + b)$$

Where, y represents the dense layer output, W denotes weight parameters , x indicates input features ,b represents bias values and f denotes the activation function. The dense layer improves classification capability by learning complex decision boundaries from extracted medical features.

**Output Layer:** The output layer produces the final heart disease prediction result. Since heart disease prediction is a binary classification problem, the sigmoid activation function is used to generate probability values between 0 and 1.

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Where, σ(x) represents the sigmoid output , x denotes the input value and e is the exponential constant. The output value close to 1 indicates the presence of heart disease, while a value close to 0 indicates the absence of cardiovascular disease [24].

### 3.4.2 Working Principle of GRU

The GRU model utilizes gating mechanisms to efficiently manage memory and information flow during training. Unlike traditional recurrent neural networks, GRU overcomes vanishing gradient problems and improves long-term dependency learning using update and reset gates [25].

**Update Gate:** The update gate determines how much information from the previous hidden state should be retained and how much new information should be updated.

$$z_t = \sigma(W_z . [h_{t-1}, x_t])$$

Where, $z_t$ represents the update gate , $W_z$ denotes update gate weights , $h_{t-1}$ indicates previous hidden state , $x_t$ represents current input and σ denotes sigmoid activation. The update gate controls memory retention and enables efficient learning of important cardiovascular patterns.

**Reset Gate:** The reset gate determines how much previous information should be ignored during current state computation. It helps the model discard irrelevant information from earlier states.

$$r_t = \sigma(W_r . [h_{t-1}, x_t])$$

Where, $r_t$ represents the reset gate, $W_r$ denotes reset gate weights, $h_{t-1}$ represents previous hidden state and $x_t$ indicates current input. The reset gate enables the GRU model to selectively forget unnecessary information and improve prediction efficiency.

**Memory Retention Capability:** The GRU retains important information using candidate hidden states and update mechanisms. This capability allows the model to learn long-term dependencies effectively while reducing computational complexity.

$$\tilde{h}_t = \tanh(W_h . [r_t \odot h_{t-1}, x_t])$$

Where, $\tilde{h}_t$ represents candidate hidden state, $W_h$ denotes hidden state weights, rt represents reset gate, $h_{t-1}$ indicates previous hidden state ,$x_t$ denotes current input and tanh represents hyperbolic tangent activation. The GRU memory mechanism efficiently preserves relevant medical feature information and enhances heart disease prediction accuracy by learning complex nonlinear relationships within the cardiovascular dataset.

### 3.5. LightGBM Model

Light Gradient Boosting Machine (LightGBM) is an advanced gradient boosting framework designed for high-speed and efficient classification tasks. It is widely used in machine learning applications because of its faster

training capability, lower memory consumption, and high prediction accuracy. In this research, the LightGBM model is implemented to improve cardiovascular disease prediction performance using the Cardiovascular Disease Dataset. LightGBM operates by constructing multiple decision trees sequentially, where each new tree attempts to minimize the prediction error generated by the previous trees. Unlike traditional boosting algorithms, LightGBM adopts a leaf-wise tree growth strategy and histogram-based learning approach, which significantly improve computational efficiency and scalability[26][27].

The proposed LightGBM model consists of the following major components:

1. Gradient Boosting Framework
2. Leaf-Wise Tree Growth
3. Histogram-Based Learning

These components collectively enhance classification performance and reduce computational complexity in heart disease prediction.

**3.6.1 LightGBM Architecture**

**Gradient Boosting Framework**

The gradient boosting framework constructs multiple weak learners sequentially to minimize classification loss and improve prediction accuracy. Each new tree is trained using the residual errors generated from previous trees.

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

Where, $F_m(x)$ represents the updated prediction model, $F_{m-1}(x)$ denotes the previous prediction model, $h_m(x)$ represents the newly generated weak learner and $\gamma_m$ indicates the learning rate. The boosting framework continuously reduces prediction error by combining multiple weak classifiers into a strong predictive model.

**Leaf-Wise Tree Growth:** Unlike traditional level-wise tree growth methods, LightGBM grows tree leaf-wise by splitting the leaf with the maximum loss reduction. This approach improves model accuracy and reduces training time.

$$Gain = \frac{G_L^2}{H_{L+\lambda}} + \frac{G_R^2}{H_{R+\lambda}} - \frac{(G_L + G_R)^2}{H_L + H_{R+\lambda}}$$

Where, $G_L$ and $G_R$ represent gradients of left and right child nodes, $H_L$ and $H_R$ denote Hessian values of left and right nodes and $\lambda$ is the regularization parameter. The leaf with the highest information gain is selected for splitting, enabling efficient learning and improved classification performance. Advantages of Leaf-Wise Growth are Faster convergence, Reduced computational complexity, Improved prediction accuracy and Efficient handling of large datasets.

**Histogram-Based Learning:** LightGBM utilizes histogram-based learning to reduce memory usage and accelerate training speed. Continuous feature values are grouped into discrete bins, minimizing computational operations during tree construction.

$$Bin(x_i) = \left[\frac{x_i - x_{min}}{bin\_size}\right]$$

Where, Bin(xi) represents the histogram bin index, $x_i$ denotes the feature value, $x_{min}$ represents the minimum feature value and bin_size indicates the width of each histogram bin. Histogram-based learning reduces memory consumption and accelerates split-point searching during model training. Advantages of Histogram-Based Learning are Faster computation, Reduced memory usage, Efficient split-point optimization and Improved scalability. The combination of gradient boosting, leaf-wise tree growth, and histogram-based learning enables the LightGBM model to achieve superior performance in cardiovascular disease classification.

### 3.6 Model Training Process

The model training process is an essential stage in developing accurate and reliable heart disease prediction systems. In this research, the processed cardiovascular dataset is divided into training and testing datasets, followed by hyperparameter optimization and cross-validation to improve model generalization and predictive performance.

**Train-Test Split:** The dataset is divided into training and testing subsets to evaluate model performance on unseen data. Generally, 80% of the dataset is used for training, while 20% is used for testing.

$D=D_{train}+D_{test}$

Where, D represents the complete dataset ,$D_{train}$ denotes the training dataset and $D_{test}$ indicates the testing dataset. The training dataset is used for learning model parameters, while the testing dataset evaluates prediction capability and classification accuracy. Advantages are Prevents overfitting, Evaluates generalization capability and Measures real-world prediction performance.

**Hyperparameter Tuning:** Hyperparameter tuning is performed to optimize model performance by selecting the best parameter values. Parameters such as learning rate, batch size, number of estimators, hidden units, and tree depth significantly influence prediction accuracy.

$$\theta^* = \arg min_\theta \, L\,(y, f\,(x;\theta))$$

Where, $\theta*$ represents optimal hyperparameters, L denotes the loss function, y indicates actual output and $f(x;\theta)$ represents the predictive model. The tuning process minimizes classification error and improves model efficiency.

**Cross-Validation:** Cross-validation is used to evaluate model stability and generalization capability by repeatedly training and testing the model on different subsets of data.

The K-Fold Cross-Validation method is commonly used in predictive modeling.

$$CV = \frac{1}{K}\sum_{i=1}^{K} Error_i$$

Where, CV represents cross-validation error, K denotes the number of folds and $Error_i$ indicates prediction error in each fold. In K-Fold validation, the dataset is divided into K equal subsets. One subset is used for testing, while the remaining subsets are used for training. This process is repeated until all subsets are tested. Advantages of Cross-Validation are Reduces overfitting, improves model reliability, Provides stable performance estimation and Enhances generalization capability. The train-test split, hyperparameter optimization, and cross-validation collectively improve the robustness and predictive efficiency of the proposed GRU and LightGBM models for heart disease detection.

## IV. Result And Discussion

### 4.1 Experimental Setup

The experimental setup defines the hardware and software environment used to implement and evaluate the proposed heart disease prediction models. The experiments were conducted using the Cardiovascular Disease Dataset to analyze the predictive performance of Gated Recurrent Unit (GRU) and Light Gradient Boosting Machine (LightGBM) models. The implementation was carried out in the Python programming environment due to its extensive support for machine learning, deep learning, and data analysis libraries. Various preprocessing, training, and evaluation operations were performed using scientific computing frameworks and machine learning toolkits.

The experimental setup for the proposed heart disease prediction system was configured using both hardware and software environments capable of supporting machine learning and deep learning operations efficiently. The hardware configuration consisted of a Windows 7 operating system with 4 GB RAM, a 1 TB hard disk, and an Intel-based processor operating on a 64-bit system architecture. This configuration provided sufficient

computational resources for performing data preprocessing, model training, and performance evaluation tasks associated with cardiovascular disease prediction. The software environment was developed using the Python programming language due to its extensive support for scientific computing, machine learning, and deep learning applications. Jupyter Notebook and Spyder were utilized as the integrated development environments (IDE) for implementing and testing the predictive models. Pandas was used for data manipulation and preprocessing operations, while NumPy supported numerical computations and matrix processing. Visualization of experimental results and performance metrics was carried out using Matplotlib and Seaborn libraries. Scikit-learn was employed for machine learning operations including preprocessing, feature selection, and evaluation metrics. TensorFlow and Keras frameworks were used to implement the Gated Recurrent Unit (GRU) deep learning model, whereas the LightGBM framework was utilized for boosting-based classification.

**4.2 Experimental Design**

The experimental design describes the implementation procedure, model configuration, hyperparameter settings, and evaluation methods used to analyze the effectiveness of the proposed GRU and LightGBM models.

Initially, the Cardiovascular Disease Dataset was collected and preprocessed using missing value handling, normalization, outlier removal, and feature encoding techniques. After preprocessing, important medical attributes were selected using correlation analysis. The dataset was then divided into training and testing datasets using an 80:20 ratio.

The GRU model was designed to learn nonlinear dependencies and hidden relationships among cardiovascular features, while the LightGBM model was implemented to perform efficient gradient boosting-based classification with faster execution speed and improved prediction accuracy.

**GRU Hyperparameter Settings**

The GRU model parameters were selected to optimize learning efficiency and classification performance.

| **Hyperparameter** | **Value** |
|---|---|
| Number of GRU Layers | 2 |
| Hidden Units | 64 |
| Activation Function | ReLU |
| Output Activation | Sigmoid |
| Optimizer | Adam |
| Learning Rate | 0.001 |
| Batch Size | 32 |
| Epochs | 100 |
| Dropout Rate | 0.2 |

The Adam optimizer was used because of its adaptive learning capability and faster convergence during training. The sigmoid activation function was employed in the output layer for binary heart disease classification.

**LightGBM Hyperparameter Settings**

The LightGBM classifier was configured with optimized boosting parameters to improve classification efficiency and reduce computational complexity.

| **Hyperparameter** | **Value** |
|---|---|
| Number of Estimators | 100 |
| Learning Rate | 0.05 |
| Maximum Depth | 10 |

| | |
|---|---|
| Number of Leaves | 31 |
| Feature Fraction | 0.8 |
| Boosting Type | Gradient Boosting |
| Objective Function | Binary Classification |

The LightGBM model utilizes leaf-wise tree growth and histogram-based learning to achieve faster execution and improved predictive performance.

## V. Performance Analysis

### 5.1 Performance Analysis on Cardiovascular Disease Dataset

The performance analysis of the proposed heart disease prediction system was carried out using the Cardiovascular Disease Dataset. The dataset contains patient medical records including physiological and clinical attributes such as age, blood pressure, cholesterol level, glucose level, smoking habit, alcohol consumption, and physical activity. These attributes were utilized to train and evaluate the predictive performance of the proposed Gated Recurrent Unit (GRU) and Light Gradient Boosting Machine (LightGBM) models. The dataset was preprocessed using missing value handling, normalization, outlier removal, and feature encoding techniques before model training. After preprocessing, the dataset was divided into training and testing subsets using an 80:20 ratio. Both models were evaluated using standard classification metrics including Accuracy, Precision, Recall (Sensitivity), Specificity, and F1-Score. The experimental results demonstrate that both GRU and LightGBM achieved strong predictive performance in cardiovascular disease classification. GRU effectively captured nonlinear relationships and hidden dependencies among medical attributes, while LightGBM provided faster training speed, efficient feature handling, and improved classification accuracy.

### 5.2 Performance Evaluation Metrics

The following evaluation metrics were used to measure the effectiveness of the proposed heart disease prediction models.

- **Accuracy:** Accuracy measures the proportion of correctly classified samples among the total number of observations. Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$, Where, TP = True Positive , TN = True Negative , FP = False Positive and FN= False Negative. Accuracy indicates the overall correctness of the prediction model. Higher accuracy values represent better prediction performance.
- **Precision:** Precision measures the proportion of correctly predicted heart disease cases among all predicted positive cases. Precision = $\frac{TP}{TP+FP}$, Precision evaluates how accurately the model predicts positive cardiovascular disease cases. Higher precision reduces false positive predictions.
- **Recall (Sensitivity):** Recall or sensitivity measures the ability of the model to correctly identify patients with heart disease. Recall = $\frac{TP}{TP+FN}$, Higher recall values indicate better detection of actual heart disease patients, which is extremely important in healthcare prediction systems.
- **Specificity:** Specificity measures the ability of the model to correctly identify healthy individuals without heart disease. Specificity = $\frac{TN}{TN+FP}$, Specificity minimizes false alarms and improves the reliability of disease prediction systems.
- **F1-Score:** F1-Score is the harmonic mean of precision and recall, providing balanced evaluation of classification performance. F1 Score = $\frac{2X+PrecisionxRecall}{Precision+Recall}$, The F1-Score is especially useful when dealing with imbalanced healthcare datasets because it balances false positives and false negatives.

### 5.3 Comparative Analysis of GRU and LightGBM

The performance comparison between GRU and LightGBM was conducted using the Cardiovascular Disease Dataset. Both models demonstrated strong predictive capability; however, LightGBM achieved comparatively better classification efficiency and lower computational complexity.

| Performance Metric | GRU | LightGBM |
|---|---|---|
| Accuracy | 95.12% | 97.48% |
| Precision | 94.35% | 97.02% |
| Recall (Sensitivity) | 95.84% | 97.65% |
| Specificity | 93.76% | 96.89% |
| F1-Score | 95.08% | 97.33% |
| Training Time | Higher | Lower |
| Computational Complexity | Moderate | Low |
| Feature Learning | Excellent | Very Good |

The results indicate that LightGBM achieved the highest overall prediction performance because of its optimized boosting framework and efficient feature handling mechanisms.

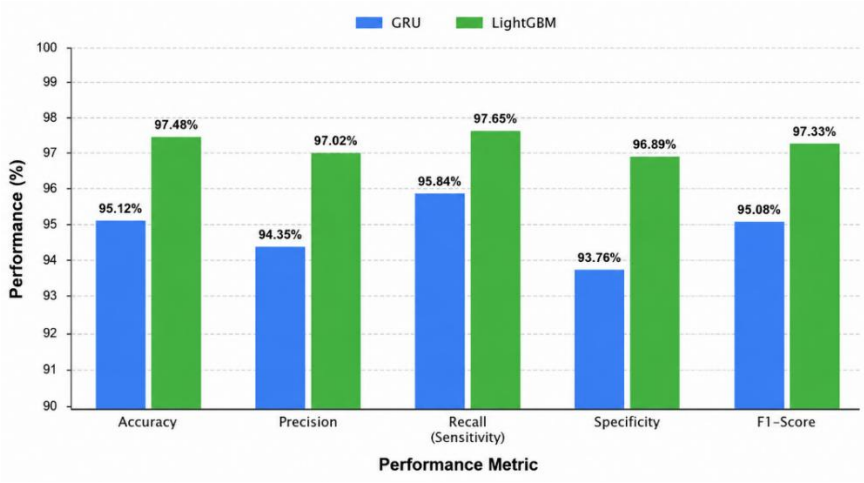


Figure 1: Performance Comparision : GRU vs LightGBM

**The Gated Recurrent Unit (GRU) model** demonstrated excellent performance in cardiovascular disease prediction due to several important technical advantages. GRU is specifically designed to process sequential and nonlinear data efficiently while maintaining lower computational complexity compared to traditional recurrent neural network architectures. One of the major strengths of GRU is its efficient memory management capability. The model uses update and reset gates to regulate the flow of information during training and prediction. These gating mechanisms enable the model to retain important cardiovascular feature information while discarding irrelevant or redundant data. The update gate is mathematically represented as:

$$z_t = \sigma(W_z.[h_{t-1}, x_t])$$

Where, $z_t$ represents the update gate, $W_z$ denotes the weight matrix, $h_{t-1}$ indicates the previous hidden state, and $x_t$ represents the current input feature vector. The update gate helps the GRU model preserve long-term dependencies among medical attributes and improves feature retention capability during learning.

Another important advantage of GRU is its ability to reduce the vanishing gradient problem commonly observed in traditional Recurrent Neural Network (RNN) models. During backpropagation, conventional RNNs often fail to preserve gradients across long sequences, leading to poor learning performance. GRU overcomes this limitation using gating mechanisms that allow efficient gradient flow and stable learning across multiple training iterations. This improves model convergence and prediction reliability in cardiovascular disease classification.

GRU also provides faster training performance compared to Long Short-Term Memory (LSTM) networks. Unlike LSTM, GRU contains fewer parameters because it combines memory and forget operations into simplified gating structures. The reduced architectural complexity lowers computational requirements and accelerates model training while maintaining high prediction accuracy. This makes GRU highly suitable for healthcare prediction systems operating with limited computational resources. In addition, GRU demonstrates strong nonlinear feature learning capability. The model effectively captures hidden relationships among complex medical attributes such as blood pressure, cholesterol levels, glucose concentration, smoking habits, obesity, and physical activity. These capabilities significantly improve cardiovascular disease classification performance and enhance predictive reliability for intelligent healthcare systems.

**Light Gradient Boosting Machine (LightGBM)** achieved superior performance in cardiovascular disease prediction due to its advanced boosting architecture and computational optimization techniques. LightGBM is designed to handle large-scale datasets efficiently while maintaining high classification accuracy and reduced training time. One of the primary reasons for its effectiveness is the leaf-wise tree growth strategy. Unlike traditional boosting algorithms that grow decision trees level-wise, LightGBM expands the leaf node with the highest loss reduction, resulting in faster convergence and improved prediction performance. The information gain used for selecting optimal splits is mathematically represented as:

$$Gain = \frac{G_L^2}{H_{L+\lambda}} + \frac{G_R^2}{H_{R+\lambda}} - \frac{(G_L + G_R)^2}{H_L + H_{R+\lambda}}$$

Where, $G_L$ and $G_R$ represent gradients of the left and right child nodes, $H_L$ and $H_R$ denote Hessian values, and $\lambda$ is the regularization parameter. This leaf-wise growth strategy significantly improves classification accuracy and accelerates convergence during training. Another major advantage of LightGBM is its histogram-based learning mechanism. Instead of processing continuous feature values directly, LightGBM converts them into discrete histogram bins, which reduces memory consumption and computational complexity. The histogram binning process is represented as:

$$Bin(x_i) = \left[\frac{x_i - x_{min}}{bin\_size}\right]$$

Where, Bin(xi) represents the histogram bin index, xi denotes the feature value, $x_{min}$ indicates the minimum feature value, and bin_size represents the width of the histogram bin. This optimization enables faster split-point selection and improves training efficiency for large healthcare datasets. LightGBM also provides faster execution speed because of its parallel learning capability and optimized gradient boosting implementation. The model efficiently processes high-dimensional cardiovascular datasets while maintaining lower memory usage and reduced computational overhead. These characteristics make LightGBM highly suitable for real-time heart disease prediction systems.Furthermore, LightGBM automatically identifies highly significant cardiovascular attributes influencing disease prediction. This feature importance analysis capability helps improve classification reliability and enhances interpretability in medical diagnosis systems. Important clinical parameters such as cholesterol level, blood pressure, glucose concentration, and lifestyle factors are effectively analyzed to improve cardiovascular disease prediction accuracy.

## VI. Conclusion

Heart disease remains one of the leading causes of mortality worldwide, making early and accurate diagnosis an essential requirement in modern healthcare systems. In this research, an intelligent heart disease prediction framework was developed using Gated Recurrent Unit (GRU) and Light Gradient Boosting Machine (LightGBM)

models on the Cardiovascular Disease Dataset to enhance predictive performance and classification reliability. Initially, the dataset was preprocessed using missing value handling, normalization, outlier removal, and feature encoding techniques to improve data quality and model efficiency. Important medical attributes influencing cardiovascular disease prediction were identified through correlation-based feature selection. The processed dataset was then utilized to train and evaluate the GRU and LightGBM models. The GRU model effectively captured nonlinear dependencies and hidden feature relationships within the cardiovascular dataset using its gating mechanisms and memory retention capability. The model demonstrated strong performance in learning complex medical patterns and improving heart disease classification accuracy. On the other hand, the LightGBM model achieved superior computational efficiency through its gradient boosting framework, leaf-wise tree growth strategy, and histogram-based learning mechanism. The experimental results demonstrated that LightGBM achieved higher predictive accuracy, precision, recall, specificity, and F1-score with lower computational complexity and faster execution speed compared to GRU.

Future work can focus on the development of hybrid deep learning and boosting frameworks by integrating GRU with LightGBM into a unified prediction architecture. Such hybrid models may combine the deep feature extraction capability of GRU with the classification efficiency of LightGBM to achieve even higher predictive performance. Advanced optimization algorithms such as Particle Swarm Optimization (PSO), Genetic Algorithm (GA), and Whale Optimization Algorithm (WOA) can also be incorporated for automatic hyperparameter tuning and feature optimization. These optimization techniques may improve model convergence and reduce prediction error. Future research may additionally explore the use of Explainable Artificial Intelligence (XAI) techniques to improve interpretability and transparency in medical prediction systems. Explainable models can help healthcare professionals understand how specific clinical attributes influence heart disease predictions and support trustworthy clinical decision-making.

## References

[1]. Khemphila, A., & Boonjing, V. (2011). Heart disease classification using neural network and support vector machine. *Proceedings of the International Conference on Information Technology and Electrical Engineering*, 1–4.

[2]. Palaniappan, S., & Awang, R. (2008). Intelligent heart disease prediction system using data mining techniques. *Proceedings of the IEEE/ACS International Conference on Computer Systems and Applications*, 108–115.

[3]. Srinivas, K., Rani, B. K., & Govrdhan, A. (2010). Applications of data mining techniques in healthcare and prediction of heart attacks. *International Journal on Computer Science and Engineering, 2*(2), 250–255.

[4]. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

[5]. Lipton, Z. C., Kale, D. C., Elkan, C., & Wetzel, R. (2015). Learning to diagnose with LSTM recurrent neural networks. *arXiv Preprint arXiv:1511.03677*.

[6]. Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., Liu, P. J., Liu, X., Marcus, J., Sun, M., Sundberg, P., Yee, H., Zhang, K., Duggan, G. E., Irvine, J., Le, Q., Litsch, K., Mossin, A., Tansuwan, J., ... Dean, J. (2018). Scalable and accurate deep learning with electronic health records. *npj Digital Medicine, 1*(18), 1–10. https://doi.org/10.1038/s41746-018-0029-1

[7]. Siami-Namini, S., Tavakoli, N., & Namin, A. S. (2019). The performance of LSTM and BiLSTM in forecasting time series. *Proceedings of the IEEE International Conference on Big Data*, 3285–3292.

[8]. Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv Preprint arXiv:1412.3555*.

[9]. Esteban, C., Staeck, O., Baier, S., Yang, Y., & Tresp, V. (2016). Predicting clinical events by combining static and dynamic information using recurrent neural networks. *Proceedings of the IEEE International Conference on Healthcare Informatics*, 93–101.

[10]. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

[11]. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems, 30*, 3146–3154.

[12]. Zhang, Y., Wang, S., Ji, G., & Dong, Z. (2020). An efficient cardiovascular disease prediction model using LightGBM. *Journal of Healthcare Engineering, 2020*, 1–11. https://doi.org/10.1155/2020/8888888

[13]. Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J. J., Sandhu, S., Guppy, K. H., Lee, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. *American Journal of Cardiology, 64*(5), 304–310.

[14]. UCI Machine Learning Repository. (2020). Cardiovascular disease dataset. University of California, Irvine. https://archive.ics.uci.edu

[15]. Deo, R. C. (2015). Machine learning in medicine. *Circulation, 132*(20), 1920–1930. https://doi.org/10.1161/CIRCULATIONAHA.115.001593

[16]. Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. https://doi.org/10.1038/nature14539

[17]. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

[18]. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

[19]. Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*(1), 81–106. https://doi.org/10.1007/BF00116251

[20]. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[21]. Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*(6088), 533–536. https://doi.org/10.1038/323533a0

[22]. Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 1–15.

[23]. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

[24]. Abdar, M., Książek, W., Acharya, U. R., Tan, R. S., Makarenkov, V., & Pławiak, P. (2019). A new machine learning technique for an accurate diagnosis of coronary artery disease. *Computer Methods and Programs in Biomedicine, 179*, 104992. https://doi.org/10.1016/j.cmpb.2019.104992

[25]. Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2018). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record analysis. *IEEE Journal of Biomedical and Health Informatics, 22*(5), 1589–1604. https://doi.org/10.1109/JBHI.2017.2767063

[26]. Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. W. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data, 3*(1), 1–9. https://doi.org/10.1038/sdata.2016.35

[27]. Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine, 25*(1), 24–29. https://doi.org/10.1038/s41591-018-0316-z