SSRG International Journal of Electrical and Electronics Engineering
ISSN: 2348-8379/ https://doi.org/10.14445/23488379/IJEEE-V11I12P135

Volume 11 Issue 12, 386-398, December 2024
© 2024 Seventh Sense Research Group®

Original Article

Improved Stochastic Gradient Descent-Decision Tree (ISGD-DT) Framework for Intelligent Heart Disease Prediction

Bollapalli Althaph[1], Nagendra Panini Challa[2]

[1,2]*School of Computer Science and Engineering (SCOPE), VIT-AP University, Andhra Pradesh, India.*

[2]*Corresponding Author : nagendra.challa@vitap.ac.in*

Received: 20 October 2024 Revised: 21 November 2024 Accepted: 19 December 2024 Published: 31 December 2024

Abstract *- This research presents an innovative system architecture for heart disease prediction that integrates Improved Stochastic Gradient Descent (ISGD) with a Decision Tree (DT) classifier. The ISGD-DT model addresses challenges in existing predictive models, such as imbalanced datasets, limited generalizability, and suboptimal accuracy, by leveraging hierarchical layers, graph databases, and decision trees for robust classification outcomes. Validated using benchmark datasets from the UCI Machine Learning Repository, including the Cleveland and Hungarian heart disease datasets, the model demonstrates superior performance with accuracy rates of 93.17%, 88.39%, and 96.29% across different datasets. These results highlight the model's reliability and robustness, making it a valuable tool for improving predictive modeling in healthcare. This research underscores the potential of combining advanced optimization techniques and classification algorithms to enhance the accuracy and applicability of medical prognostics.*

Keywords *- Heart illness prediction, Decision Tree, Improved Stochastic Gradient Descent, Deep Learning.*

1. Introduction

The heart is vital for human health, playing a critical role in circulating oxygenated blood and regulating key bodily functions. Cardiovascular Diseases (CVDs), including Coronary Heart Disease (CHD), remain the leading cause of death globally, accounting for approximately 17.9 million fatalities annually, as reported by the World Health Organization (WHO). The prevalence and mortality rates associated with CVDs underscore the urgent need for effective diagnostic and predictive solutions [1]. Despite advancements in Artificial Intelligence (AI) and Machine Learning (ML), current heart disease prediction models face significant challenges, including imbalanced datasets, limited generalizability, and suboptimal accuracy. These limitations hinder their clinical applicability, leading to biased forecasts and reduced reliability in real-world scenarios. Several risk factors contribute to the development of cardiovascular diseases, including high BP, obesity, abnormal lipid profiles, diabetes, smoking, lack of physical activity, excessive alcohol consumption, and high cholesterol levels. The WHO projects that cardiovascular illnesses will remain a leading cause of death well into the future, presenting a significant threat to human health, potentially even beyond 2030. In this context, ML offers significant potential for transforming healthcare, as noted by KSL Prasanna et al. [2]. ML's advanced data processing capabilities exceed human ability, leading to innovative solutions for complex healthcare challenges. In recent years, Artificial Intelligence (AI) applications, particularly ML, have been increasingly used to identify cardiovascular disorders with speed and precision. Despite progress, there is still a pressing need to refine predictive prototypes and address research gaps, such as the challenge of imbalanced datasets, which can lead to biased forecasts. Researchers have explored various methodologies, including NN and DL techniques, to develop hybrid models that enhance forecast accuracy [3–12]. While these studies provide valuable insights, the differences in datasets, models, and outcomes highlight the complexity of predicting cardiovascular diseases.

Although improvements have been made, further research is essential to advance existing models and enhance overall prediction accuracy. The growing use of DL in this field emphasizes the ongoing need for continued exploration to improve the reliability and applicability of prediction models, ultimately leading to more effective clinical interventions and improved patient care. In our study, datasets from the UCI repository, including the Cleveland and Hungarian heart disease datasets, were utilized to further explore this critical area of research. Existing predictive models for heart disease often face significant challenges, including:

- Imbalanced datasets: Many models struggle to provide accurate predictions due to the uneven distribution of disease and non-disease cases.
- Limited generalizability: Predictive accuracy varies across different datasets and patient populations, limiting clinical applicability.
- Integration issues: Inefficient use of diverse data types, such as structured clinical records and unstructured patient data, hampers comprehensive analysis.
- Suboptimal accuracy: Current models do not achieve the reliability needed for effective clinical decision-making.

2. Literature Survey

Siddiqui S. et al. [13] examine the application of ANN and BN in classifying diabetes and cardiac illness. Alic, B. et al. [13] utilize the Levenberg-Marquardt learning method, a type of multilayer feed-forward neural network, as an ANN method to test the hypothesis that it can enhance the precision of diabetes and heart illness diagnosis by providing more reliable statistical data. Ozcan M et al. [14] suggested a novel cardiac illness forecast model based on random forest. This model outperformed the benchmark multivariate regression ideal and other models like CART, NB, Bagged Trees, and AdaBoost. Researchers designed the model to assess the 3-year risk of heart illness. The study employed the random forest algorithm on a substantial dataset to assess the likelihood of cardiovascular illness in eastern China. Kasbe, T et al. [15] GD is a technique that commonly optimizes multiple loss functions, particularly linear functions. This context has utilized stochastic gradient descent to address the root-finding aspect of cardiovascular disorders. SGD selects random samples for each iteration using a batch, representing the sample size instead of the entire dataset.

Each iteration computes the gradient using specific batches. Using GDS for diagnosing cardiovascular illness yielded a relatively high accuracy rate of 84.39%. Li, Y., Sperrin, M et al. [16] deem the identification of cardiovascular illness crucial for life-saving purposes. The DCD-DEML approach, which employs back-propagation, obtained a diagnostic precision of 92.45% in identifying cardiovascular disease. This accuracy is superior to the DCD Mamdani Fuzzy Inference System and the DCD ANN. Hashi, E. K et al. [17] Medical practitioners desire a comprehensive diagnostic tool for accurately identifying cardiac failure based on the provided information. The implemented fuzzy expert system comprised three main components: fuzzification, a rule base, and defuzzification. This system was built utilizing MATLAB's Fuzzy Logic Toolbox and operated on the Mamdani Fuzzy Inference System framework. The measurements of precision and sensitivity yielded high values, specifically 94.50% and 90.19%, respectively. Patro, S. P. et al. [18] found that auto-prognosis significantly improved the accuracy of cardiovascular risk forecasts compared to other high-performing systems. This method was established using data collected from over 400,000 members of the UK Biobank, with 450 parameters recorded for each individual. The approach was developed to investigate new cardiovascular risk variables without any preconceived biases systematically. An evaluation was conducted to compare the therapeutic validity of the auto-prognosis model with the classic Framingham model. The auto-prognosis algorithm accurately forecasted outcomes for 3,357 out of 4,801 cardiovascular patients. Poornima V. et al. [19] have extensively researched machine learning systems for forecasting cardiac illness. Algorithms such as Naïve Bayes (NB), Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF) were applied to the dataset from the UCI Machine Learning Repository. The analysis revealed that the RF system achieved a maximum precision of 90.16% in forecasting heart illness. Buchan K. et al. [20] advanced a hybrid classifier to forecast heart illness. They selected the attributes using the orthogonal local preservation forecast method. Artificial Neural Networks (ANN) carry out categorization. The neural network architecture consisted of four neurons in the input layer, one hundred in the hidden layer, and five in the output layer. The connection weights between neurons ranged from -10 to 10. The Group Search Optimization (GSO) technique and the Levenberg-Marquardt (LM) algorithm were applied to optimise the network. The final weights were selected from the two sets generated by the LM and GSO methods. The authors could verify the results' precision by utilizing three datasets—one from Cleveland, Hungary, and Switzerland. The structure achieved an accuracy rate of 98% on the Hungarian dataset, and on the Switzerland dataset, it reached 87%. On the Cleveland dataset, it reached 94%.

Budholiya K. et al. [21] predicted the occurrence of illness by considering risk factors such as elevated cholesterol levels, a lack of physical exercise, hypertension, and an unhealthy dietary pattern. The novelists used computerized medical accounts, which consist of unstructured data. Mdhaffar A. et al. [22] utilized NLP and ML methods to generate forecasts from unstructured data. The novelists utilized the i2b2 Heart Illness Risk Issues Challenge dataset, which consisted of 296 patient records with diabetes. Kevin Challa, N. P. et al. [14] Heart illness and diabetes share certain risk factors that accelerate diabetes's progression. This was a challenge for the academics. For NLP purposes, the novelists used Apache cTAKES. Using the data obtained from cTAKES, the model was developed using Principal Component Analysis (PCA) and mutual information for feature selection. Following the feature selection process, Maximum Entropy (MaxEnt), Support Vector Machine (SVM), and Naïve Bayes (NB) classifiers were employed to perform the classification task, achieving an F1-Score of 77.4%. Louridi N. et al. [23] proposed a framework for implementing a stacked collective prototype. A stacked prototype was constructed using XGBoost, Gradient Boosting (GB), and Random Forest (RF) classifiers, with dimensionality reduction executed using Particle Swarm

Optimization (PSO). The algorithm reached a precision of 93.55% on the Statlog dataset, 86.49% on the Cleveland dataset, and 91.18% on the Hungarian data collection. Bashir S. et al. [24] created an HRF using a linear technique to classify heart illness. A DT selects a feature based on the entropy value. The Cleveland data collection system achieved a precision rate of 88.7%. The researchers used the Mean, Mode, KNN, and MICE algorithms Tomar, D et al. [25] Olaniyi E. O et al. [26] to fill in the missing data. Additionally, the dataset underwent class balancing. The stacking algorithm obtained an accuracy of 95.83%. Jothi, K et al. [27] created a predictive algorithm for heart illness utilizing the collective mechanism. The researchers conducted experiments on five datasets. A 1 or 0 class label designates the presence or absence of illness respectively. The inter-quantile range approach was utilized for outlier detection. Various classification techniques were applied, including Support Vector Machines (SVM), Decision Trees (DT), Naïve Bayes (NB), and memory-based classifiers. To enhance accuracy, the outputs of these classifiers were combined using the majority vote method. This approach achieved precision levels of 86.81% on the Cleveland dataset, 81.15% on the SPECTF records, 82.35% on the SPECT dataset, 86.21% on the Eric records, and 88.26% on the Statlog records. Manogaran G. et al. [28] developed a system based on least-squares twin SVMs, employing F-scores for attribute selection. Experiments using the Statlog dataset yielded an accuracy of 85.59%. Deepa, N. et al. [29] proposed a technique utilizing Multilayer Perceptrons (MLP) and SVM, with the MLP trained using the back-propagation method. The MLP had a 0.32 knowledge rate.

Table 1. Comparison table

S.no	Datasets	Limitations	Methods	Accuracy
1	Cleveland	The model's limitations include not incorporating patients' medical and social factors, excluding unstructured data, and lacking comprehensive data for broader generalization.	Regression Tree (CART) algorithm, a supervised machine learning.	87%,
2	Cleveland	The paper highlights the complexities of heart disease diagnosis, emphasizing the impact of human biases, the limitations of expert judgment, and the need for advanced data mining to improve decision-making in healthcare.	The research uses the AllPossible-MV algorithm for missing value imputation, the C4.5 decision tree for rule generation, and hill climbing for rule subset optimization, and it evaluates performance with 10-fold cross-validation.	86.3% accuracy in testing and 87.3% in training.
3	Heart disease dataset	The research highlights the importance of data quality in decision-making, identifies limitations in existing heart disease analysis methods, and emphasizes the need for hybrid technologies to improve results and address current constraints.	Naïve Bayes, BO-SVM, KNN, and SSA-NN	93.3%
4	UCI Cardiac Dataset	The paper highlights limitations in data source customization, the need for further exploration of feature combinations, concerns about scalability and resource requirements in real-time prediction systems, and the lack of analysis of dataset characteristics' impact on prediction performance.	The paper uses a DTRF classifier with SGB optimization for heart disease prediction, employing data preprocessing, bootstrapped training, and performance evaluation based on precision, recall, F1 score, and accuracy.	precision of 86%, recall of 86%, F1-score of 85%, and accuracy of 96%
5	UCI Machine Repository	The research with a limited dataset of 1025 instances highlights concerns about model accuracy and misdiagnosis, suggesting future work with larger datasets and more attributes to improve heart disease diagnosis.	Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Naïve Bayes (NB), Artificial Neural Network (ANN), Random Forest (RF), and Gradient Descent Optimization (GDO)	Accuracy of 98.54%, sensitivity of 99.43%, and precision of 97.76%
6	Heart disease dataset	The research faced limitations due to reliance on a single dataset, lack of traditional confidence intervals, and impractical bootstrap sampling, with future studies encouraged to use broader datasets for enhanced robustness.	SMOTE, ADASYN, SMOTE-Tomek, and SMOTE-ENN, logistic regression, decision trees, random forest, gradient boosting, XGBoost, CatBoost, and Artificial Neural Networks (ANNs)	recall rate of 88% and an AUC of 82%

The SVM reached a precision rate of 87.5%, while the MLP achieved a precision rate of 85%. Nawaz, M. S. Aet al. [30] used decision trees and KNN classifiers to resolve cardiac illness forecasting. The KNN technique achieved a precision of 67%, while the DT algorithm achieved a precision of 81%.

3. Proposed Methodology

Researchers extensively use the UCI Heart Illness dataset and the Kaggle Heart Illness dataset, which combine data from Statlog, Cleveland, and Hungary, to predict heart illness. Both offer essential cardiovascular health attributes, including age, BP, cholesterol levels, and types of chest discomfort, which facilitate creating and assessing machine learning models for forecasting heart illness risk. Figure 1 illustrates the ISGD-DT model's procedures.

The process began with three preparation stages: format conversion, data transformation, and data normalization. Following these preparatory steps, sample selection and 10-fold cross-validation were performed. Data classification was conducted using the ISGD-DT model, which combines Improved Stochastic Gradient Descent (SGD) with Decision Trees (DT) for effective categorization. The ISGD-DT model was evaluated using a benchmark dataset, with results analyzed across multiple epochs.

3.1. Improved Stochastic Gradient Descent (ISGD)

Improved Stochastic Gradient Descent (ISGD) is a widely used optimization technique in ML and DL. ISGD is an efficient optimization technique that necessitates real-time monitoring and utilizes memory storage. For example, consider a collection that contains several instances. Typically, improved stochastic gradient descent processes more observations for each iteration. The calculation of variables can be enhanced to facilitate rapid evaluation of web learning with new observations by processing individual data points simultaneously in the Improved Stochastic Gradient Descent (ISGD) method. A random input a and a scalar output b, represented as a pair (a, b), compose each z sample.

When the correct answer is y, the loss function Q (b, b) evaluates the detection cost. It chooses a family F of functions ggkk (a) with a weight vector k. The function gg E can minimize the loss function RR (c, k) = P (ggkk (a), b) on instances. The laws of nature are invariably determined based on observations derived from a sample c1... cn, independent of the unknown distribution eQ(C).

$$d(g) = \int m(g(a), b) e\, Q(C) F_o(g) = \frac{1}{o} \sum_{j=1}^{o} \ell(g(a_j), b_j) \qquad (1)$$

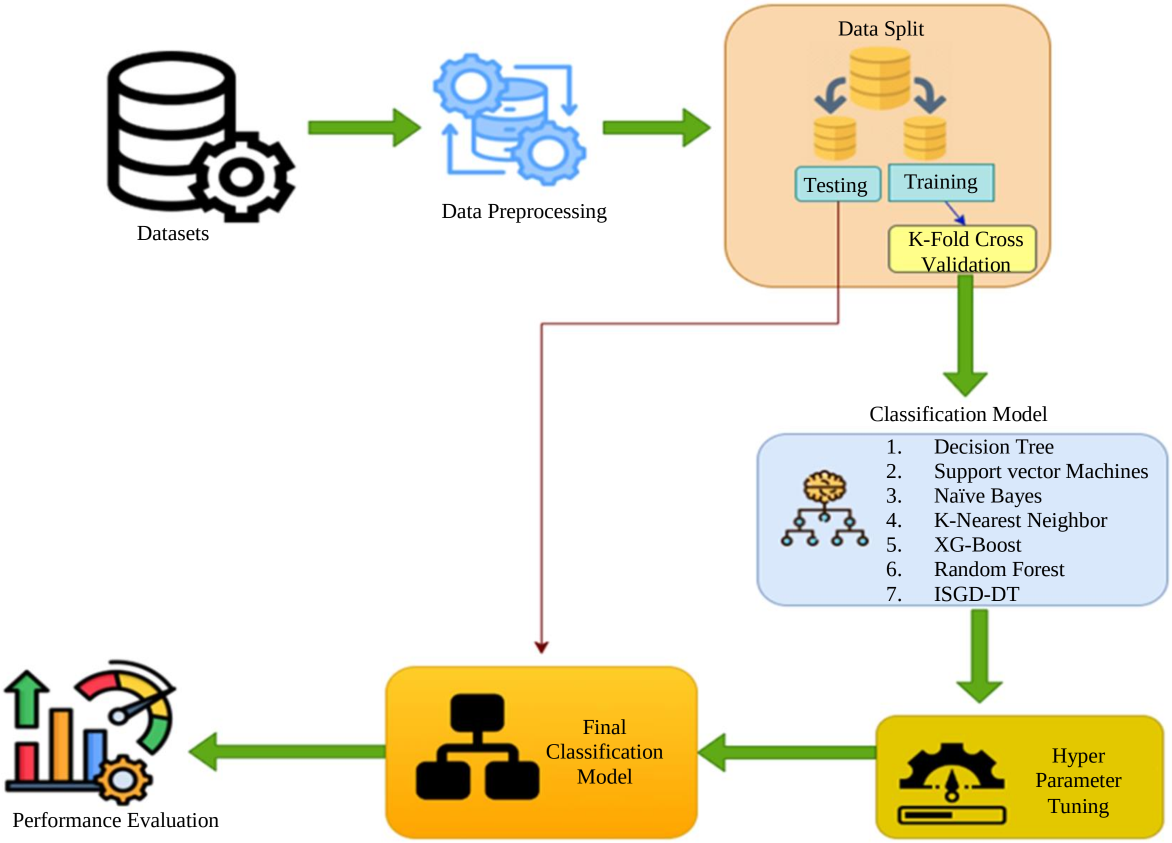


Fig. 1 Proposed ISGD-DT method

The empirical risk is Fo (gg). The targeted risk E (gg) calculates the expected generalizing operation for each subsequent event. Statistical learning theory suggests that constraining a selected family F leads to decreased empirical risk rather than projected danger. Typically, Gradient Descent (GD) is used to reduce the empirical risk En (ggkk). Using a gradient, every iteration incrementally raises the weight k.

$$\mathrm{k_{s+1}} = \mathrm{k_s} - \gamma \frac{1}{\mathrm{o}} \sum_{\mathrm{j=1}}^{\mathrm{o}} \nabla_{\mathrm{k}} \mathrm{R}(\mathrm{c_j}\,, \mathrm{k_s}) \qquad (2)$$

The symbol $\gamma\gamma$ represents a carefully selected learning rate. When functions exhibit sufficient regularity, achieving linear convergence is possible if the initial estimate k0 is near the optimal value and the learning rate $\gamma\gamma$ is low. Log t, which represents the remaining error, denotes linear convergence.

$$\mathrm{k_{s+1}} = \mathrm{k_s} - \tau_s \frac{1}{\mathrm{o}} \sum_{\mathrm{j=1}}^{\mathrm{o}} \nabla_{\mathrm{k}} \mathrm{R}(\mathrm{z_j}\,, \mathrm{k_s}) \qquad (3)$$

The Newton model commonly implements the Second-order Gradient Descent (2GD) technique. By maximizing the regularity concerns, 2GD achieves quadratic merging when the provided value k0 is close to the optimal value. When the cost is quadratic, and the transformation matrix is correct, the method achieves its highest value after just one iteration. Otherwise, these returns will be satisfactory if they are sufficiently smooth $loglog\ \rho \sim t$. The ISGD method is a substantial popularization strategy. Each subsequent iteration computes the gradient by replacing the gradient descent of Fo (ggkk) with a singular value Cs.

$$\mathrm{k_{s+1}} = \mathrm{k_s} - \gamma_s \nabla_{\mathrm{k}} \mathrm{R}\big(\mathrm{z_j}\,, \mathrm{k_s}\big) \qquad (4)$$

The ISGD model deliberately processes instances, recalling them from earlier iterations. The models are taken from ground truth spreading, and Improved Stochastic Gradient Descent (ISGD) is optimized accordingly. Table 1 illustrates an Improved Stochastic Gradient Descent (ISGD) approach for traditional machine learning methods. Primarily used for Perceptron, Adaline, and K-means mapping. Conventional optimization approaches were employed to configure the Support Vector Machine (SVM) and Lasso models. In both cases, a hyper-parameter controls the regularization term. RR $_{\mathrm{SVM}}$ and RR $_{\mathrm{lasso}}$. Due to RR means being a non-convex function, the K-means algorithm converges to a local minimum. The predicted update rule includes 2GD learning parameters to guarantee rapid convergence. When you use the Improved Stochastic Gradient Descent (ISGD) algorithm on these parameters and ensure they are positive, you get solutions with fewer non-zero elements. The stochastic approximation literature extensively researches the convergence of Improved Stochastic Gradient Descent (ISGD). When outcomes converge, they tend to have reduced learning values to meet the restrictions of the equations $\sum s\,\gamma\gamma^2 < \infty$ and$\sum s\,\gamma\gamma_t < \infty$. The Robbins-Sigmund theorem enables the achievement of nearly certain convergence despite challenging circumstances, such as when the loss function is non-smooth. The noisy approximation of a positive gradient slowed down ISGD's convergence speed. Reducing the learning value gradually minimizes the variance of a parameter estimate wt. If learning rates decline, it takes longer for the variable estimate, k $_{\mathrm{s}}$, to reach the best answer. When the Hessian matrix of a cost function has conditionally positive eigenvalues, the fastest convergence speed can be achieved by using learning rates $\gamma\gamma t$ ∼s−1. The rate at which the desire to remain error-free decreases is proportional to time, where D(Б) ∼ s −1. Theoretical convergence values are commonly detected ∼ s−1. The rate at which the desire to remain error-free decreases is proportional to time, where D (Б) ∼ s −1. Theoretical convergence morals are commonly detected. The purposes of D (ρ) ∼ s−1/2 generally converge. Convergence is observed experimentally during the final stage of job optimization. The factor is not considered significant, t, as the optimization procedure ends before achieving the necessary solution.

3.2. Second-Order Improved Stochastic Gradient Descent (2ISGD)

Second-order Improved Stochastic Gradient Descent (2ISGD)uses a positive definite matrix s to get close to the inverse of the Hessian matrix and add the gradients.

$$\mathrm{k_{s+1}} = \mathrm{k_s} - \gamma_s \tau_s \nabla_{\mathrm{k}} \mathrm{R}\big(\mathrm{z_j}\,, \mathrm{k_s}\big) \qquad (5)$$

The variation unexpectedly fails to reduce stochastic noise and does not improve $\mathrm{w_t}$. As constants rise, the anticipated residual error decreases, following a D(ρ) ∼ s−1 pattern at its most optimal. DGS's optimisation model becomes progressively slower compared to the general batch approach. Several fields, including biomedical research, commerce, criminology, ecology, engineering, and healthcare, use Decision Trees (DT) as a classification method. Decision Trees (DT) are classified as generalized linear methodology. Generalized linear methods evaluate the regression function when dealing with binary parameters, while decision tree approaches apply to continuous variables. These methods compare the dependent parameter y with many predictor values to determine the required value. A DT is a discriminative classifier that directly studies the mapping from input x to output y by creating the following probability (b | a). DT's parametric technique is outlined here. The modest function can represent the decision tree as a sigmoid function according to Equation (6), among other models.

$$\sigma(a) = \frac{1}{1+e^{-z}} \qquad (6)$$

It is referred to as a loss function, which quantifies the difference between the predicted outcomes and the actual values in a model, the 0–1 losses for a specific method.

$$Loss\,\frac{0}{1}(z) = \{\, \begin{matrix} 1, if\ z < 0 \\ 0, otherwise \end{matrix} \qquad (7)$$

Let $y\varepsilon\{-1, 1\}$ and z=b.k$^{\mathrm{S}}$a. If y and wTx have the same sign, z is positive; otherwise, it is negative.

$$(b = -\tfrac{1}{x}) = \frac{1}{1+exp(k_o+\sum_{j=1}^{e} k_j a_j)} \quad (8)$$

$$q(b = \tfrac{1}{x}) = 1 - q(b = -\tfrac{1}{x}) \quad (9)$$

The primary function of decision trees is to minimize k, resulting in a decrease in the maximum value of $0 - 1$ loss compared to training themes.

$$min \sum_{j=1}^{o} l\tfrac{0}{1}(b^j . k^S . a^j) \quad (10)$$

$$k = [k_1, k_2, k_3, \ldots\ldots\ldots\ldots\ldots . k_n] \leftarrow$$

$$arg_k max\ \Pi_w Q(b^{(w)} | a^w, k) \quad (11)$$

Graphing the 0/1 loss function transforms the regression approach into a logistic function. The values range from 0 to 1, while z varies from -∞ to +∞.

$$l_{\log}(z) = \log(1 + e^{-z}) \quad (12)$$

Moreover, the gradient descent rule is applied to weight k. The primary goal of constructing decision trees is to manage continuous features and effectively handle nominal and missing values. Its illustrates the distribution of logistic losses that commonly occur. Regularization incorporated into the learning process helps prevent overfitting by filtering out the irregular features in the dataset. R1 and R2 mostly achieve regularization, leading to sparsity in reducing complexity. A decision tree algorithm that focuses on regularization learns a mapping (k) that reduces the logistic loss on the training data by adding a regularization term. Regularization in decision trees involves utilizing a higher figure of likelihood functions, as defined in Equation (13).

$$\underset{k}{min} \sum_{j=1}^{o} l_{\log}\left(b^{(j)} . k^S . b^j\right) + \lambda \parallel k \parallel \overset{2}{2} \quad (13)$$

Equation (13 consists of the training log-loss function and the model struggle. The λ derived from model complexity serves as a regularization parameter. It calculates the w variables that need to be increased. By utilizing Equation (13) as a cost function, the outcome of a suggestion may reduce over-fitting. Choosing a large value for w results in smoothing and can cause under-fitting. By regularly applying $L1$ regularization, many techniques lead to reducing variables to 0, resulting in a sparse parameter vector in the simulation results.

3.3. Experimental Findings and Performance Assessment

An extensive analysis was conducted on Datasets 1 and 2 to confirm the effectiveness of the proposed technique. False Positive Rate, False Negative Rate, sensitivity, specificity, accuracy, and F-score are the measures utilized for analyzing the results. Table 3 displays the presentation metrics used to evaluate the outcomes of the suggested examples.

Table 2. Comparison table structure

Model	Datasets	Accuracy	Precision	F-Measure
ISGD-DT	Dataset-1	86.61	91.52	93.17
	Dataset-2	81.82	81.82	88.39
Random Forest (RF)	Dataset-1	86.11	85.98	58.66
	Dataset-2	75.53	75.12	63.55
Support Vector Machines (SVM)	Dataset-1	85.16	85.16	55.86
	Dataset-2	79.10	78.22	67.55
Naïve Bayes (NB)	Dataset-1	86.44	86.44	55.88
	Dataset-2	78.22	78.22	68.10
K-Nearest Neighbor (KNN)	Dataset-1	86.60	86.60	57.39
	Dataset-2	75.88	75.88	64.14
XGBoost	Dataset-1	85.96	85.96	55.99
	Dataset-2	75.98	75.98	63.66

Table 3. Metrics for evaluating performance

S. No.	Variables	Notation
1	FPR	$\frac{FP}{FP + FN}$
2	FNR	$\frac{FN}{FN + FP}$
3	Sensitivity	$\frac{TP}{TP + FN}$
4	Specificity	$\frac{TN}{TN + FP}$
5	Precision	$\frac{TP + TN}{TP + TN + FP + FN}$

6	F-Score	$\frac{2TP}{2TP+FP+FN}$
7	Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$

Table 4. Confusion matrix generated using ISGD-based DNN on dataset 1

Specialists	Epochus-100		Epochus-200		Epochus-300		Epochus-400		Epochus-500	
	Heart Illness	Non-Heart Illness	Heart Illness	Non-Heart Illness	Heart Illness	Non-Heart Illness	Heart Illness	Non-Heart Illness	Heart Illness	Non-Heart Illness
Heart Illness	2840	20	2840	20	2840	20	2840	20	2840	20
Non-Heart Illness	483	10	483	10	483	10	483	10	483	10

3.4. Analysis of Results on Dataset 1

Testing was conducted on Dataset 1 by varying the number of epochs in increments of 100, 200, 300, 400, and 500, as presented in Table 4. The results remained consistent across 100, 200, 300, 400, and 500 epochs. 2,860 incidents were correctly identified as churn, with no instances classified as non-heart illnesses. The classification outcomes are derived from the confusion matrix and organized according to various metrics, including false positive rate, false negative rate, sensitivity, specificity, precision, and F-score. It is recommended that these evaluation parameters be used to assess the model's performance comprehensively; false positive and negative rates have low values in this case. Simultaneously, the sensitivity, specificity, precision, and F-score rates must be elevated. Table 5 and Figure 2 display the categorization results achieved with varying numbers of epochs. The analysis of the table and figure reveals that the false negative rate is 15.09, the sensitivity is 87.06%, the accuracy is 86.55%, and the F-score value is 93.17% when 100 epochs are considered. Similarly, The same outcomes are observed when the classifiers are run for 200, 300, 400, and 500 epochs.

Table 5. Performance of various iterations on dataset 1

No. of Runs	FPR	FNR	Sensitivity (%)	Specificity (%)	Precision	F-Score
Epchos-100	36.02	15.09	87.06	67.09	87.56	93.17
Epchos-200	35.04	16.09	86.14	67.19	83.06	93.17
Epchos-300	34.82	15.09	87.06	68.13	88.15	93.17
Epchos-400	34.23	15.09	81.12	68.95	89.86	93.17
Epchos-500	33.22	17.05	87.06	67.15	86.55	93.17

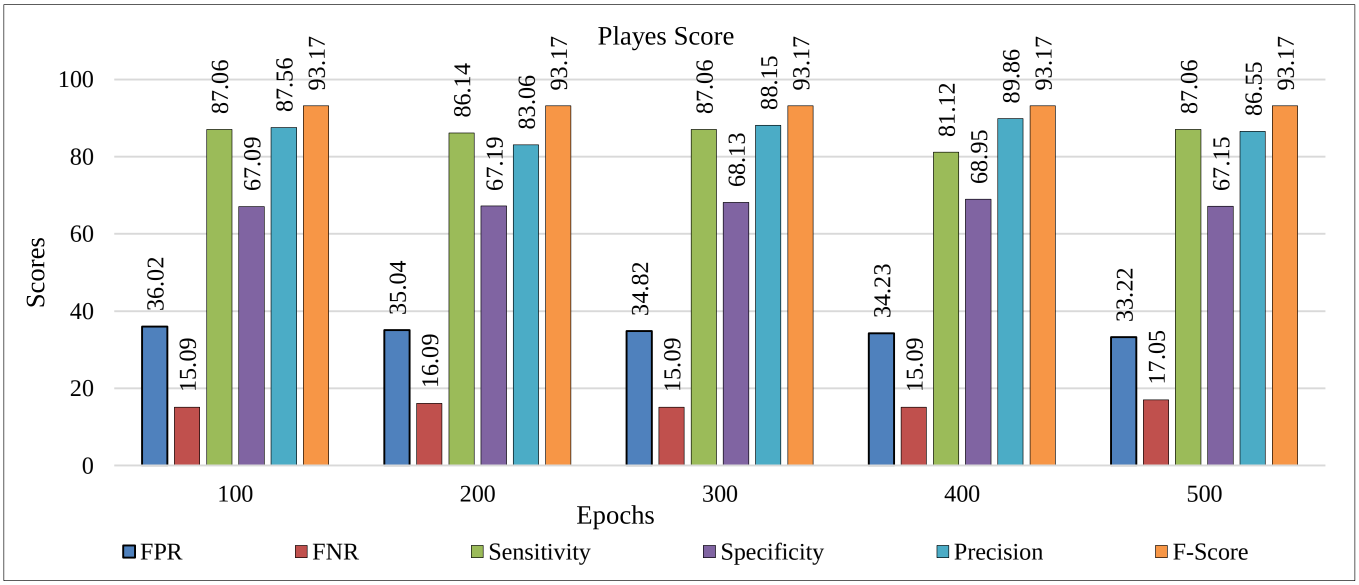


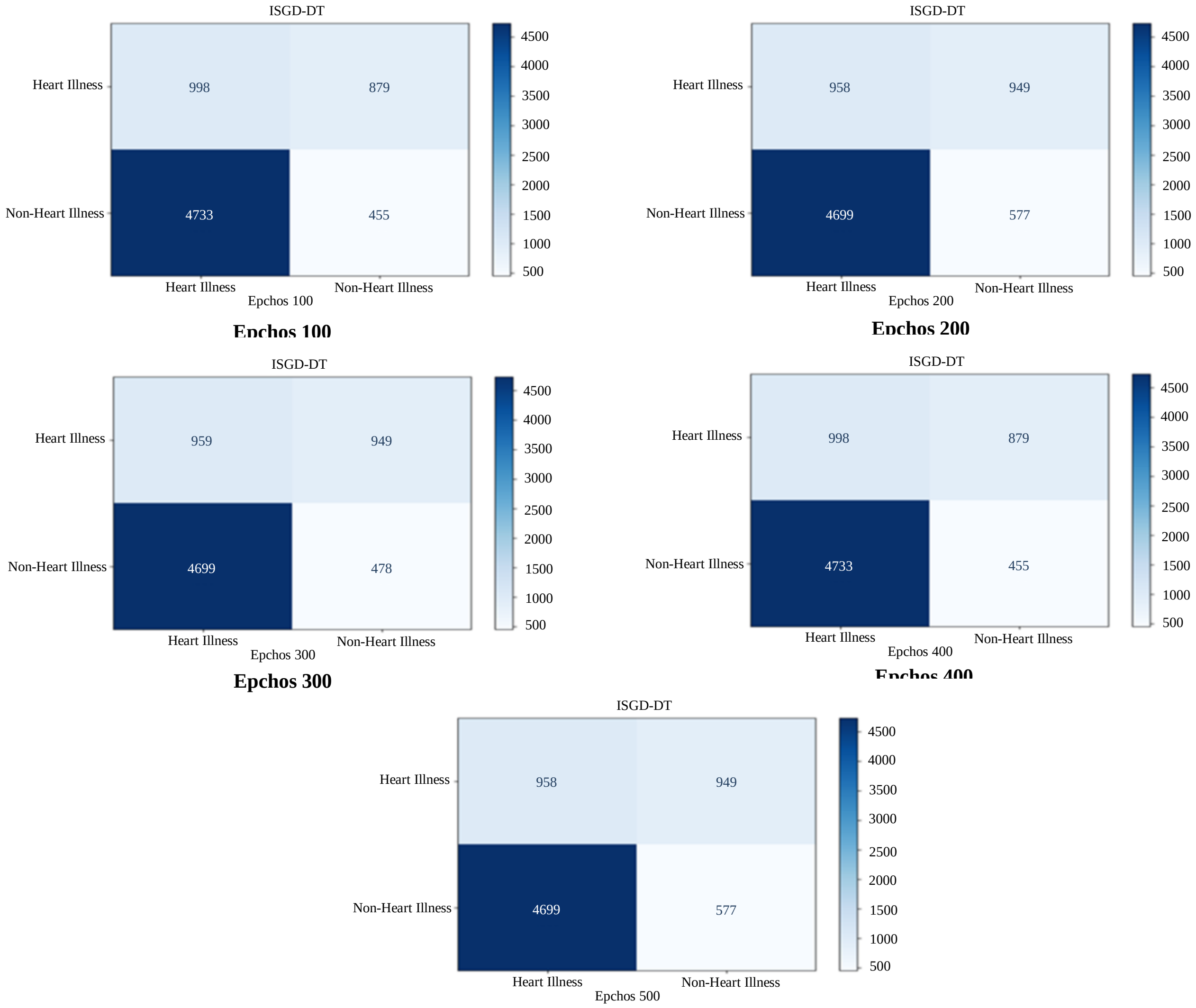
Fig. 2 Performance evaluations for dataset 1

3.5. Analyzing the Results from Dataset 2

An experiment was conducted on the applied dataset, varying the number of epochs in increments of 100, 200, 300, 400, and 500, as shown in Table 6. For 100 epochs, 4,733 cases were correctly identified as churns, and 879 instances were accurately classified as non-diseases. With 200 epochs, 4,699 cases were correctly identified as churns, while 949 instances were correctly labeled as non-diseases. At 300 epochs, the same results were observed, with 4,699 churn cases and 949 non-disease instances correctly identified. For 400 epochs, 4,706 churn cases and 963 non-disease instances were accurately classified. Finally, after 500 epochs, 4,718 instances were correctly categorized as Diseases, and 966 cases were accurately labeled as non-diseases. The results indicate that as the number of epochs increases, the complexity of the classifier's presentation also rises.

Table 6. Confusion matrix generated using ISGD-DT for dataset 2

Specialists	**Epochus-100**		**Epochus-200**		**Epochus-300**		**Epochus-400**		**Epochus-500**	
	Heart Illness	**Non-Heart Illness**	**Heart Illness**	**Non-Heart Illness**	**Heart Illness**	**Non-Heart Illness**	**Heart Illness**	**Non-Heart Disease**	**Heart Illness**	**Non-Heart Illness**
Heart Illness	998	879	958	949	959	949	998	879	958	949
Non-Heart Illness	4733	455	4699	577	4699	478	4733	455	4699	577



Epchos 100

Epchos 200

Epchos 300

Epchos 400

Epchos 500

Fig. 3 Confusion matrix generated using ISGD-DT for dataset 2

Table 7. Performance of various iterations on dataset 2

No. of Runs	FPR	FNR	Sensitivity (%)	Specificity (%)	Accuracy	F-Score
Epchos-100	35.02	18.10	83.06	66.89	80.25	87.84
Epchos-200	35.04	17.28	84.41	66.91	80.15	87.55
Epchos-300	34.82	17.91	84.66	67.88	81.44	87.71
Epchos-400	34.23	17.26	84.58	67.91	81.87	88.58
Epchos-500	33.22	17.12	84.87	67.53	81.53	88.77

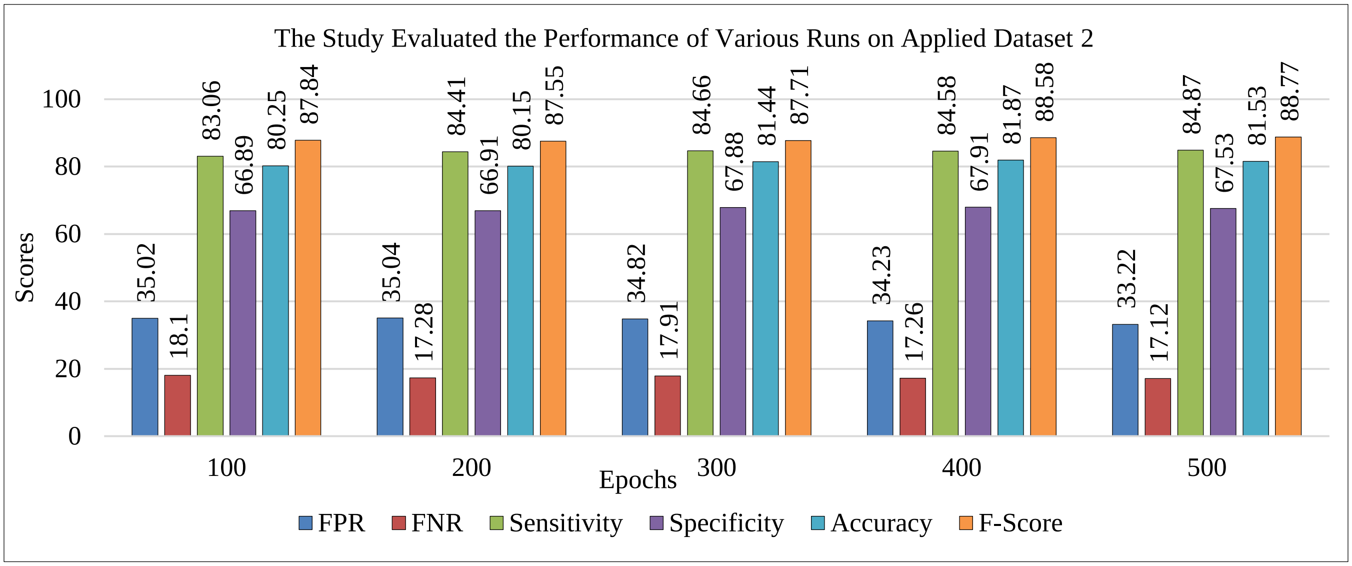


Fig. 4 Performance evaluation on dataset 2

Table 7 and Figure 4 display the categorization results achieved with varying numbers of epochs. The table reveals that the false positive rate is 35.02, the false negative rate is 18.10, the sensitivity is 83.06%, the specificity is 66.89%, the precision is 80.25%, and the F-score is 87.84% when 100 epochs are considered. The following performance metrics were observed after 200 epochs: The model achieved a false positive rate of 35.04, a false negative rate of 17.28, a sensitivity of 84.41%, a specificity of 66.91%, a precision of 80.15%, and an F-score of 87.55%. Similarly, the model achieves a false positive rate of 34.82, a false negative rate of 17.91, a sensitivity of 84.66%, a specificity of 67.88%, a precision of 81.44%, an F-score of 87.71%, and 300 epochs. Over 400 epochs, the model demonstrates a false positive rate of 34.23, a false negative rate of 17.26, a sensitivity of 84.58%, a specificity of 67.91%, a precision of 81.87, and an F-score of 88.58. After 500 epochs, it is noteworthy that the false positive rate is 33.22, the false negative rate is 17.12, the sensitivity is 84.87%, the specificity is 67.53, the precision is 81.53%, and the F-score is 88.77%. The provided data clearly shows that using 500 epochs leads to a peak precision of 88.77%, indicating a development in the classifier's performance as the number of epochs increases.

3.6. A Comparative Analysis of Current Approaches for Practical Datasets

Table 8 assesses prior models for datasets 1 and 2, using precision and F-measure as metrics. Next, we compare dataset 1 and conventional methodologies, focusing on accuracy and F-measure. Table 8 and Figure 5 present the analysis. The data presented in the table indicates that the current methodology achieves a notable level of precision, specifically 85.16%, and an F-measure of 55.86%. The ISGD-DT technique performs an augmented classification task, yielding an accuracy rate of 86.61% and an F-measure of 93.17%. Therefore, the aforementioned extensive experimental research confirms the effectiveness of the ISGD-DT method as a classification tool for heart illness prediction. According to the table, when applied to dataset 2, the previous models provided a superior accuracy of 79.10% and an F-measure of 67.55%.

As a result, the ISGD-DT methodology has a remarkable classification efficacy, with an accuracy rate of 81.82% and an F-measure of 88.39%. Therefore, the experimental research has shown that the ISGD-DT framework is a suitable classification technique for forecasting heart illness. This training aims to classify heart illness using the ISGD technique and the DT classifier model. The integration of ISGD and DT can lead to an effective classification. The ISGD-DT model's performance is evaluated using a benchmark dataset, with results analyzed across various epochs. This paper describes the remarkable classification performance of the model, achieving accuracy rates of 86.61%, 81.82%, and 95.63% for the three employed datasets. Furthermore, the F-measure for the aforementioned datasets is documented as 93.17%, 88.39%, and 96.29%, respectively.

Table 8. A comparative analysis between the proposed technique and existing approaches for applied datasets

Techniques	Dataset-1		Dataset-2	
	Precision (%)	F-Measure (%)	Precision (%)	F-Measure (%)
ISGD-DT	91.52	93.17	81.82	88.39
DT	85.98	58.39	75.12	62.99
SVM	85.16	55.86	79.10	67.55
NAÏVE BAYES	86.44	55.88	78.22	68.10
KNN	86.60	57.39	75.88	64.14
XG-Boost	85.96	55.99	75.98	63.66
RF	86.11	58.66	75.53	63.55

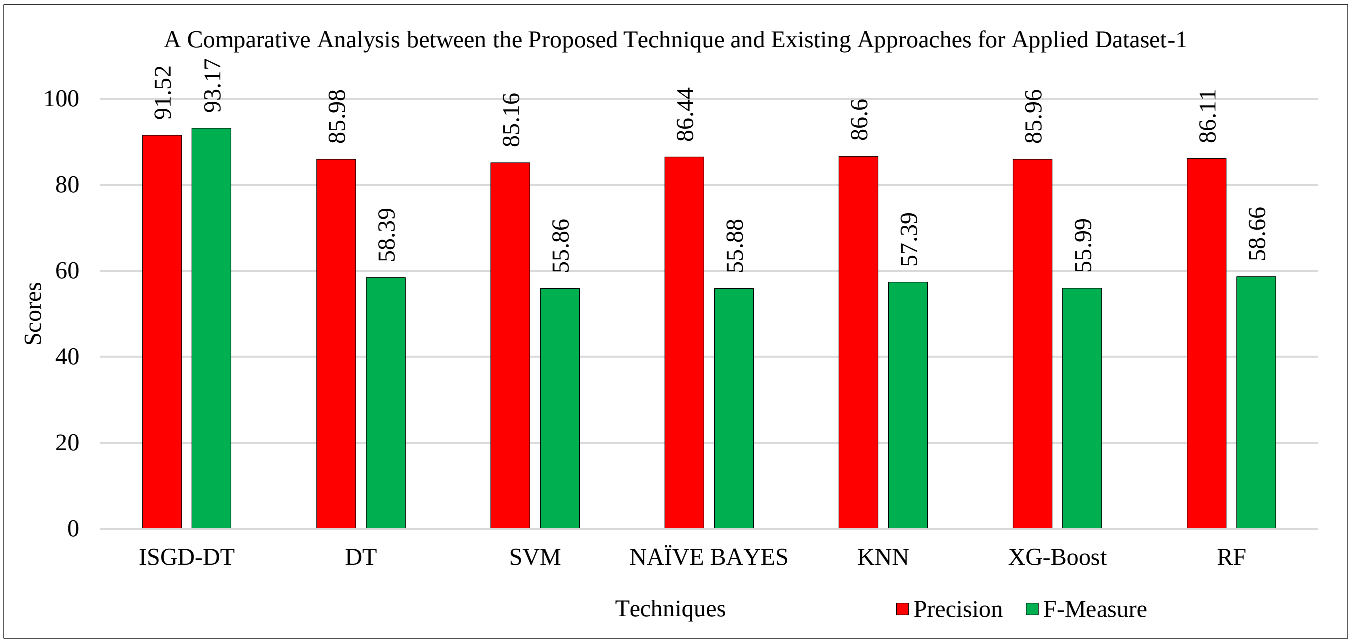


Fig. 5 Performance comparison between the ISGD-DT with existing approaches for dataset-1

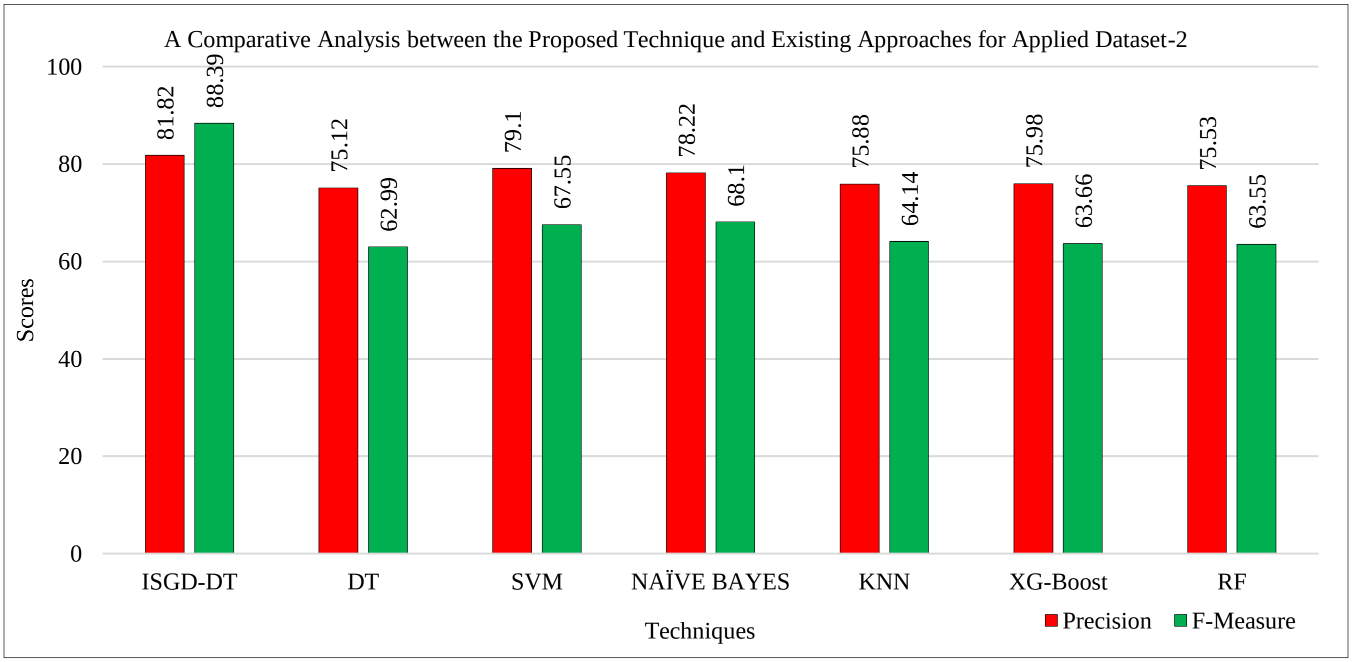


Fig. 6 Performance comparison between the ISGD-DT with existing approaches for dataset-2

4. Limitations and Future Scope

This study introduces the Improved Stochastic Gradient Descent-Decision Tree (ISGD-DT) framework, designed to overcome the limitations of existing methodologies by:

Leveraging advanced optimization techniques and robust classification algorithms. Utilizing hierarchical architecture, graph databases, and decision trees for enhanced data integration and predictive accuracy. Traditional models such as Decision Trees (DT), Naïve Bayes (NB), and Support Vector Machines (SVM) have achieved accuracy levels ranging between 75-86% on heart disease datasets. However, these models often lack robustness and fail to address issues like dataset imbalance and diverse data integration. The ISGD-DT framework improves upon these methodologies by integrating stochastic gradient descent for dynamic optimization and leveraging decision trees for accurate classification. With accuracy rates of 86.61%, 81.82%, and 95.63% across three benchmark datasets and F-measures of 93.17%, 88.39%, and 96.29%, it demonstrates significant improvements over traditional and hybrid approaches.

4.1. Dependency on Dataset Quality

The model's performance is contingent on the quality and representativeness of the datasets used. The inclusion of diverse and larger datasets from different demographics could improve generalizability. Computational Requirements: The ISGD-DT framework involves computationally intensive processes, especially during model training with large datasets, which may limit scalability in resource-constrained environments. Handling of Unstructured Data: While the model integrates structured data effectively, it cannot process unstructured data such as clinical notes, which could provide additional insights.

4.2. Real-Time Applicability

The framework's real-time prediction capabilities in clinical settings require further validation to ensure reliability under varying conditions. Incorporating Real-World Data:

Future studies could integrate Electronic Health Records (EHRs) and wearable device data to enhance the model's predictive accuracy and applicability. Unstructured Data Integration: Developing methods to incorporate unstructured data, such as textual clinical notes effectively, could provide a more comprehensive analysis.

4.3. Optimization for Real-Time Use

Improving computational efficiency and exploring cloud-based solutions can make the model suitable for real-time clinical deployment.

4.4. Personalized Prediction Models

Extending the framework to create personalized predictions based on individual patient characteristics could enhance clinical utility.

5. Conclusion

This study systematically explores research methodologies for predicting heart disease, explaining multi-layered system architecture. The approach employs an Improved Stochastic Gradient Descent with Decision Tree (ISGD-DT) classifier model to enhance prediction accuracy. An effective classification technique was developed by integrating information systems and graph databases with decision trees. The model's performance was evaluated using a benchmark dataset, with results analyzed across multiple epochs. The ISGD-DT model demonstrated strong classification capabilities, achieving accuracy rates of 86.61%, 81.82%, and 95.63% across three different datasets, along with F-measures of 93.17%, 88.39%, and 96.29%. These findings confirm the model's robustness and reliability as a predictive tool for heart disease classification.

Data Availability

https://archive.ics.uci.edu/dataset/45/heart+disease, https://www.kaggle.com/datasets/sid321axn/heart-statlog-cleveland-hungary-final

Acknowledgement

Here is a revised version of the sentence in third-person perspective:

Gratitude is extended to VIT-AP University for providing infrastructure and technical support.

Author Contributions

Bollapalli Althaph: writing, analysis, and preparing a draft manuscript. Methodology and reviewing the manuscript. Dr. Nagendra Panini Challa is responsible for verifying and validating the manuscript.

References

[1] Cardiovascular Diseases, Health Topics, World Health Organization, 2019. [Online]. Available: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1

[2] María Teresa García-Ordás et al., "Heart Disease Risk Prediction Using Deep Learning Techniques with Feature Augmentation," *Multimedia Tools and Applications*, vol. 82, no. 20, pp. 31759-31773, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[3] P. Ramprakash et al., "Heart Disease Prediction Using Deep Neural Network," *International Conference on Inventive Computation Technologies (ICICT)*, Coimbatore, India, pp. 666-670, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[4] S. Sambath Kumar, and M. Nandhini, “Entropy Slicing Extraction and Transfer Learning Classification for Early Diagnosis of Alzheimer Diseases with sMRI,” *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 17, no. 2, pp. 1-22, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[5] Ivo Sérgio Guimarães Brites et al., “Machine Learning and Iot Applied to Cardiovascular Diseases Identification Through Heart Sounds: A Literature Review,” *Informatics*, vol. 8, no. 4, pp. 1-24, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[6] Adedayo Ogunpola et al., “Machine Learning-Based Predictive Models for Detection of Cardiovascular Diseases,” *Diagnostics*, vol. 14, no. 2, pp. 1-19, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[7] Pooja Rani et al., “An Extensive Review of Machine Learning and Deep Learning Techniques on Heart Disease Classification and Prediction,” *Archives of Computational Methods in Engineering*, vol. 31, no. 6, pp. 3331-3349, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[8] Xiaofeng Yuan, Lin Li, and Yalin Wang, “Nonlinear Dynamic Soft Sensor Modeling with Supervised Long Short-Term Memory Network,” *IEEE Transactions on Industrial Informatics*, vol. 16, no. 5, pp. 3168-3176, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[9] Surenthiran Krishnan, Pritheega Magalingam, and Roslina Ibrahim, “Hybrid Deep Learning Model Using Recurrent Neural Network and Gated Recurrent Unit for Heart Disease Prediction,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 6, pp. 5467-5476, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[10] Rohit Bharti et al., “Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning,” *Computational Intelligence and Neuroscience*, vol. 2021, no. 1, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[11] Li Yang et al., “Study of Cardiovascular Disease Prediction Model Based on Random Forest in Eastern China,” *Scientific Reports*, vol. 10, no. 1, pp. 1-8, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[12] Khaled Mohamad Almustafa, “Prediction of Heart Disease and Classifiers’ Sensitivity Analysis,” *BMC Bioinformatics*, vol. 21, no. 1, pp. 1-18, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[13] Shahan Yamin Siddiqui et al., “Modelling, Simulation and Optimization of Diagnosis Cardiovascular Disease Using Computational Intelligence Approaches,” *Journal of Medical Imaging and Health Informatics*, vol. 10, no. 5, pp. 1005-1022, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[14] Mert Ozcan, and Serhat Peker, “A Classification and Regression Tree Algorithm for Heart Disease Modeling and Prediction,” *Healthcare Analytics*, vol. 3, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[15] Tanmay Kasbe, and Ravi Singh Pippal, “Enhancement in Diagnosis of Coronary Artery Disease Using Fuzzy Expert System,” *International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT)*, vol. 3, no. 3, pp. 1324-1331, 2018. [Google Scholar] [Publisher Link]

[16] Yan Li et al., “Consistency of Variety of Machine Learning and Statistical Models in Predicting Clinical Risks of Individual Patients: Longitudinal Cohort Study Using Cardiovascular Disease as Exemplar,” *BMJ*, vol. 371, pp. 1-9, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[17] Emrana Kabir Hashi, and Shahid Uz Zaman, “Developing A Hyperparameter Tuning Based Machine Learning Approach of Heart Disease Prediction,” *Journal of Applied Science and Process Engineering*, vol. 7, no. 2, pp. 631-647, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[18] Sibo Prasad Patro, Gouri Sankar Nayak, and Neelamadhab Padhy, “Heart Disease Prediction by Using Novel Optimization Algorithm: A Supervised Learning Prospective,” *Informatics in Medicine Unlocked*, vol. 26, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[19] V. Poornima, and D. Gladis, “A Novel Approach for Diagnosing Heart Disease with Hybrid Classifier,” *Biomedical Research*, vol. 29, no. 11, pp. 2274-2280, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[20] Kevin Buchan, Michele Filannino, and Özlem Uzuner, “Automatic Prediction of Coronary Artery Disease from Clinical Narratives,” *Journal of Biomedical Informatics*, vol. 72, pp. 23-32, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[21] Kartik Budholiya, Shailendra Kumar Shrivastava, and Vivek Sharma, “An Optimized Xgboost Based Diagnostic System for Effective Prediction of Heart Disease,” *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 7, pp. 4514-4523, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[22] Afef Mdhaffar et al., “CEP4HFP: Complex Event Processing for Heart Failure Prediction,” *IEEE transactions on Nanobioscience*, vol. 16, no. 8, pp. 708-717, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[23] Nabaouia Louridi, Samira Douzi, and Bouabid El Ouahidi, “Machine Learning-Based Identification of Patients with A Cardiovascular Defect,” *Journal of Big Data*, vol. 8, no. 1, pp. 1-15, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[24] Saba Bashir et al., “MV5: A Clinical Decision Support Framework for Heart Disease Prediction Using Majority Vote Based Classifier Ensemble,” *Arabian Journal for Science and Engineering*, vol. 39, no. 11, pp. 7771-7783, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[25] Divya Tomar, and Sonali Agarwal, “Feature Selection Based Least Square Twin Support Vector Machine for Diagnosis of Heart Disease,” *International Journal of Bio-Science and Bio-Technology*, vol. 6, no. 2, pp. 69-82, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[26] Ebenezer Obaloluwa Olaniyi, Oyebade Kayode Oyedotun, and Khashman Adnan, "Heart Diseases Diagnosis Using Neural Networks Arbitration," *International Journal of Intelligent Systems and Applications*, vol. 7, no. 12, pp. 75-82, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[27] K. Arul Jothi et al., "WITHDRAWN: Heart Disease Prediction System Using Machine Learning," *Materialstoday: Proceedings*, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[28] Gunasekaran Manogaran, and Daphne Lopez, "Health Data Analytics Using Scalable Logistic Regression with Stochastic Gradient Descent," *International Journal of Advanced Intelligence Paradigms (IJAIP)*, vol. 10, no. (1-2), pp. 118-132, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[29] N. Deepa et al., "An AI-Based Intelligent System for Healthcare Analysis Using Ridge-Adaline Stochastic Gradient Descent Classifier" *The Journal of Supercomputing*, vol. 77, no. 2, pp. 1998-2017, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[30] Muhammad Saqib Nawaz, Bilal Shoaib, and Muhammad Adeel Ashraf, "Intelligent Cardiovascular Disease Prediction Empowered with Gradient Descent Optimization," *Heliyon*, vol. 7, no. 5, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[31] Purushottam, Kanak Saxena, and Richa Sharma, "Efficient Heart Disease Prediction System," *Procedia Computer Science*, vol. 85, pp. 962-969, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[32] Anil Pandurang Jawalkar et al., "Early Prediction of Heart Disease with Data Analysis Using Supervised Learning with Stochastic Gradient Boosting," *Journal of Engineering and Applied Science*, vol. 70, no. 1, pp. 1-18, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[33] Konstantina-Vasiliki Tompra, George Papageorgiou, and Christos Tjortjis, "Strategic Machine Learning Optimization for Cardiovascular Disease Prediction and High-Risk Patient Identification," *Algorithms*, vol. 17, no. 5, pp. 1-23, 2024. [CrossRef] [Google Scholar] [Publisher Link]