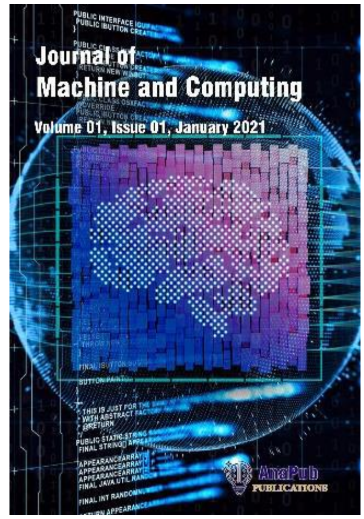
Journal Pre-proof

DI-CVD Tri-Layer CX Classifier for Secure IoT-Enabled Risk Prediction Model



Thumilvannan S and Balamanigandan R

DOI: 10.53759/7669/jmc202505124

Reference: JMC202505124

Journal: Journal of Machine and Computing.

Received 12 February 2025

Revised form 22 April 2025

Accepted 28 May 2025

Please cite this article as: Thumilvannan S and Balamanigandan R, "DI-CVD Tri-Layer CX Classifier for Secure IoT-Enabled Risk Prediction Model", Journal of Machine and Computing. (2025). Doi: https://doi.org/10.53759/7669/jmc202505124.

AnaPub
PUBLICATIONS

DI-CVD Tri-Layer CX Classifier for Secure IoT-Enabled Risk Prediction Model

S. Thumilvannan[1], R. Balamanigandan[2*]

[1,2]Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha University, SIMATS, Chennai, Tamilnadu, India.

[1]stvvannan@gmail.com, [2]balamanigandanr.sse@saveetha.com

Corresponding Author: R. Balamanigandan

ABSTRACT

This paper introduces a novel Di-CVD Tri-Layer CX Classifier, an IoT-integrated and machine learning (ML)-driven framework, to predict the individual and joint risk of diabetes (DB) and heart disease (HD). The proposed model comprises three phases: secure IoT-based data collection using Enhanced BGV encryption with Dynamic Distributed Hashing (DDH); a feature extraction (FE) phase leveraging (IGO) Information Gain Ratio and disease-specific ranking and a three-step classifier—Cm-Ro (FS) feature selection, hierarchical XGBoost classification, and synergistic prioritized risk scoring. By integrating multi-attribute features, rule-free optimization, and enhanced interoperability, the model addresses critical challenges such as heterogeneous data formats, poor feature relevance, and low interoperability in previous studies. When compared to conventional classifiers such as SVM and standard XGBoost, experimental evaluation on the NHANES dataset shows improved performance in terms of accuracy (ACC), recall (R), precision (P), and F1-score. The outcomes validate the framework's effectiveness in early, secure, and individualized risk prediction, offering substantial support for timely interventions and enhanced patient care.

Keywords: Diabetes and Heart Disease Prediction; IoT-integrated Healthcare; Machine Learning Classifier; Feature Extraction and Selection; Encrypted Health Data Processing

1. INTRODUCTION

The leading causes of death worldwide are DB and cardiovascular disease (CVD). About thirty percent of all deaths were caused by this CVD and DB [1].Ninety percent of diseases can be avoided with early detection (ED), according to a global survey. Patients must be regularly diagnosed, and their risk factors for disease must be examined. IoT has been created for that reason. In a number of ways, IoT improves patient care. IoT-connected sensors in medical devices, pressure sensors, and hospital wristbands are some of the efficient uses of this technology. [2] [3]. Health professionals can monitor patient data in a real-time (RT) environment on display due to these sensors' ability to gather and transmit data to the server.

As a result, less manual data entry and collection is needed. IoT-enabled devices can also be used by healthcare facilities to continually monitor vital signs such as blood pressure (BP), body temperature, glucose levels, and heart rates (HR) [4]. The health worker will be able to determine whether a medical emergency is present at the appropriate moment if this information is tracked and provided in real-time (RT). Death rates will drop as a result. Nonetheless, these devices are often quite heterogeneous, generating vast quantities of fitness and health data in a variety of formats. Data extraction from various medical devices is a daily challenge for hundreds of healthcare organisations, impacting medical research as well as patient care [5] [6].All of these healthcare organisations, are having a lot of difficulties in handling these massive volumes of data, mostly because they do not have an integrated system for exchanging data. Interoperability is the only way to allow systems to communicate with one another and exchange data with as many organisations as feasible. Hence, achieving high interoperability in IoT-based medical data transfer (DT) is necessary.

Because of the increase in the amount of data being collected and the need to boost the system's intelligence, ML techniques are being used in the medical industry [7] [8]. A ML model comprising multiple medical datasets will be trained utilising a variety of health care data pertaining to disease. Both patients and healthcare professionals will be associated with all of those data. Personalised treatment and behavioural modification, drug manufacturing and the identification of novel patterns that lead to new medications and treatments, clinical trial research, and smart (EHR) electronic health records are all areas in which ML can be useful.

Existing methods for risk prediction have limitations. Some of them are listed below.

- ❖ Rule-based (RB) processing of disease variables and classifying them was done previously. However, these rules lack confidence levels, and the increase in rules, in turn, increases the time for building a classifier.[14]
- ❖ Previous risk prediction models considered a number of factors such as BMI, BP Cholesterol, etc.; but they failed to consider its activity barrier such as fear of reactions from hypoglycemia.[15]

- ❖ Considering DB and HD risk prediction, most of the models do not indicate risk prediction (RP) of any particular disease but did it for overall disease; hence, individual disease RP is required.[16]
- ❖ Previous DB prediction models had difficulty predicting diabetes from the specified nine attributes; also, they were permitted for optimal feature selection (FS) by mapping, therefore, still, it remains a tedious task to identify the most accurate FS procedure for ML.
- ❖ In heart disease prediction, the classification was fed with categorical data as input hence they failed to be a multi-relational classification model and this resulted in the elimination of memory bound for future risk prediction.
- ❖ IoT machine learning models previously used for secure communication between patients and predictors, these models in the network get fragmented with respect to communication formats; this made it difficult for devices to communicate and limited the lack of interoperability.

Therefore, for predicting the risk score level in diseases, ML is used as a risk prediction tool. Prior DB and HD risk prediction algorithms used proven risk factors like age, smoking, high BP, cholesterol, and DB to predict future risk. The majority of risk prediction algorithms have identified etiological connections between these risk factors and HD and DB features. Missing values (MV) in the dataset are the major drawback that resulted in features that directly affect the accuracy of the decision-making (DM) tool. Many people who are at danger are still not recognised by these methods, and some individuals who are not at risk receive unnecessary preventative care. Thereby, improvement in these risk prediction tools is necessary.

2. LITERATURE SURVEY

Ma et al. [9] conducted a study for the prediction of cardiovascular disease patients and diabetes patients with high risk. Data on CHD patients from the eight cycles of the Health and Nutrition Examination Survey (NHANES) were gathered for this retrospective cohort analysis. Using hazard ratios (HRs) and 95% confidence intervals (CIs), competing risk models were created to assess the relationship between HR and CV death. Among CHD patients, an increased risk of CV death was linked to an HR of <70 or ≥80 bpm. However, this study considered pulse rate as heart rate, which was measured only once; also, this study does not represent any inter/intra reliability with pulse measurement; hence, the result of this study remains unsatisfactory.

Chen et al [10] presented a regression model (RM)-based Energy-adjusted score calculation. This score was computed using 24-hour food recall data; sex-specific thresholds were defined by low muscle mass and strength, and the knee extensor kinetic strength (peak force) was assessed using this value. The NHANES 1999–2002 dataset was utilised in this cross-sectional study to assess performance. Regression model results were inconsistent because a number of methods were used in this model to find the score; this resulted in low accuracy.

The possibility of detecting Type-II DB just by using age, body mass index (BMI), and glycated haemoglobin (HbA1c) was confirmed by **Thamaraimanalan et al. [11]** using a method based on K-means clustering (KMC). Based on age, BMI, and HbA1c, the NHANES dataset has been assigned to pre-established subgroups. 10-fold cross-validation (CV) was used to analyse the three variables' classification performances, and logistic regression (LR) and Cox regression analysis (CoRA) were used to evaluate the results. The model's effectiveness was reduced by type 1 DB confounding, even if those under 30 were not allowed to forecast Type 2 DB.

Using data from the NHANES database, **Zhang et al. [12]** looked into the relationship among dietary fibre intake and long-term CVD risk. The 10-year risk of CVD among individuals was predicted using the atherosclerotic-CVD score, which took into account the participants' age, sex, race, cholesterol, BP, medication use, DB status, and smoking status. The data's normality was examined using the Shapiro-Wilk test (SWT). However, due to a lack of data, this study does not examine the impact of soluble and insoluble fibre intake on the risk of CVD, which led to erroneous risk prediction conclusions.

In order to predict prediabetes in the provided dataset, **Tu et al. [13]** developed a risk prediction model. Initially, the relationships between the combined lifestyle scores and health outcomes in each cohort were measured using Cox proportional-hazards regression models. Then, a random-effects meta-analysis approach was used to pool multivariable-adjusted hazard ratios (HRs). Before analysing them, it first thought that the event CVD was incident ischaemic HD and stroke. The extent to which healthy lifestyle factors were linked to lowering the risks of CVD, cancer, and mortality in individuals with prediabetes was not demonstrated by this model, which exhibits ambiguous results.

3. PROPOSED METHODOLOGY

The proposed Di-CVD Tri-Layer CX Classifier framework is designed to provide secure, accurate, and individualized risk prediction for both diabetes (DB) and heart disease (HD) by leveraging IoT-based health data and machine learning techniques. This methodology is structured into three integrated phases: secure data acquisition, feature engineering and ranking, and a multi-stage classification model. Figure 1 shows the proposed block diagram.

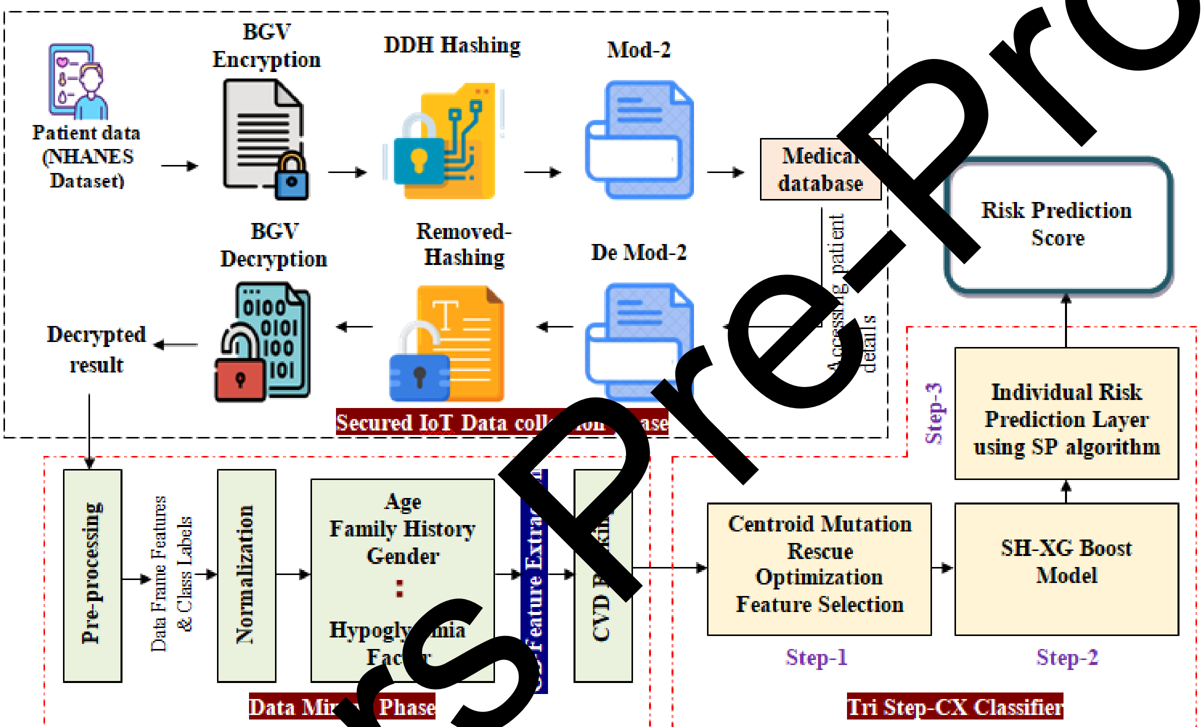


Figure 1. Block Diagram of proposed Tri-Layer CX Classifier

3.1. Secure IoT-Based Data Collection

In the first phase of the Di-CVD Tri-Layer CX Classifier framework, a network of IoT-enabled biomedical sensors such as glucose monitors, heart rate sensors, blood pressure cuffs, and wearable activity trackers is employed to gather real-time physiological and behavioral health data from patients. These devices continuously capture multi-dimensional data including blood pressure (BP), heart rate (HR), glucose levels, and physical activity metrics, all of which are crucial indicators for early detection of diabetes and cardiovascular disease. The primary challenge in this phase is not only the reliable acquisition of data but also ensuring the confidentiality, integrity, and authenticity of the transmitted data, particularly in an environment susceptible to security threats and privacy violations.

To achieve end-to-end confidentiality, the collected data is encrypted using an Enhanced Brakerski-Gentry-Vaikuntanathan (BGV) encryption scheme, which is a variant of homomorphic encryption. Homomorphic encryption allows computations to be performed on ciphertexts, enabling analysis without revealing the underlying data. The standard BGV encryption scheme operates over polynomials and allows additive and multiplicative operations to be conducted in the encrypted domain. The encryption of a message m under a secret key sk typically yields a ciphertext c, such that:

$$c = E_{sk}(m) = m + e \mod q \tag{1}$$

where:

- m is the plaintext,
- e is a small error term (to ensure semantic security),
- q is a large modulus.

The enhanced BGV used in this model introduces an additional modulus reduction step using $mod\ 2$ at the final ciphertext layer. This enhancement reduces the ciphertext space to binary levels, increasing the resistance against brute-force decryption and reverse polynomial attacks. It can be represented as:

$$c' = (c \mod 2) \mod q \tag{2}$$

This step ensures that the ciphertext remains indistinguishable even under known-ciphertext attacks, especially in low-noise environments such as wearable IoT systems.

Once encryption is complete, the data undergoes Dynamic Distributed Hashing (DDH) to ensure integrity and secure indexing. The DDH process involves the generation of unique hash values for each encrypted data packet, such that:

$$h_i = H(c_i' \| t_i) \tag{3}$$

where:

- h_i is the hash value for the ith encrypted data packet,
- c_i' is the final encrypted data,
- t_i is the timestamp, and

- $H(\cdot)$ is a secure one-way hash function like SHA-256.

This timestamp concatenation provides temporal traceability and eliminates replay attacks, while the hashing ensures data tamper resistance.

The encrypted and hashed data packets are then securely transmitted over a medical-grade communication network to a central server or cloud platform, where health practitioners perform the decryption and validation process. Upon reception, the system verifies the hash h_i for each packet by recomputing it from the received ciphertext and timestamp. If $h_i^{\text{received}} = h_i^{\text{computed}}$, the data is accepted as authentic and unaltered.

Decryption of the ciphertext c' is then performed using the private key sk, extracting the original message:

$$\hat{m} = D_{sk}(c') = m \tag{4}$$

Only if the hash verification passes is the data integrated into the patient's health record for further analysis. This secure pipeline ensures that patient data remains confidential, authentic, and unaltered from the moment of collection to its use in predictive analytics.

3.2. Data Preprocessing, Feature Engineering, and Di-CVD-Specific Ranking

Once the encrypted health data collected from IoT devices is securely transmitted and decrypted in the central server environment, the second phase begins with comprehensive data preprocessing and feature engineering. This phase aims to transform raw heterogeneous clinical data into a high-quality, structured format suitable for machine learning analysis. The first step involves the elimination of duplicate records, typically using hash-based or row-wise comparison techniques. Subsequently, missing values in key attributes such as glucose level, blood pressure (BP), and body mass index (BMI) are imputed.

After cleansing, the data undergoes normalization to reduce feature value skewness and scale all numerical attributes into a consistent range, typically [0, 1]. Once normalized, the dataset is passed through a feature selection pipeline based on Information Gain Ratio (IGR), which quantifies the relevance of each feature with respect to the class label—either No Risk (0) or Risk (1). IGR improves upon the traditional information gain by penalizing attributes with a large number of distinct values and is computed as:

$$IGR(A) = \frac{IG(A)}{H(A)} \ \ Where \ \ IG(A) = H(T) - H(T|A) \tag{5}$$

Here,

- H(T) is the entropy of the target class:

$$H(T) = -\sum_{i=1}^{n} p_i \log_2 p_i \tag{6}$$

- where p_i is the probability of class i,
- $H(T \mid A)$ is the conditional entropy given attribute A,
- $H(A)$ is the intrinsic information of attribute A.

The IGR score of each feature allows for quantitative ranking within predefined rule sets, enabling automated selection of features that reduce classification uncertainty. The output is a feature matrix where rows represent patients and columns correspond to highly ranked attributes.

Key features considered for diabetes (DB) risk prediction include family history, age, gender, race and ethnicity, height, weight, blood glucose level, and HbA1c levels. A novel and domain-specific attribute, termed the Hypoglycemia Factor (HF), is also integrated into the model. HF accounts for the risk associated with excessive insulin intake or diabetes medication, which may lower glucose levels to dangerous thresholds. A sample formula for estimating HF could be:

$$HF = \frac{\text{Insulin Dosage} \times \text{Medication Sensitivity Factor}}{Glucose\ Baseline\ Level} \tag{7}$$

This factor improves the accuracy of DB risk classification by incorporating behavioral and pharmacological risk components.

For heart disease (HD), in addition to standard features such as cholesterol, BP, BMI, and ECG anomalies, the methodology introduces a Di-CVD Priority Ranking Score (DPRS) to emphasize physical activity levels, as inactivity is a leading contributor to cardiovascular conditions. The DPRS assigns higher weights to features with stronger empirical associations to HD using a relevance function:

$$\mathrm{DPRS_i} = \alpha . f_i + \beta . g_i \tag{8}$$

Where:

- f_i = physical activity relevance index for feature i,
- g_i = statistical significance of feature i from prior clinical studies,
- α and β = empirically tuned weights based on correlation analysis.

The combination of HF for diabetes and DPRS for heart disease enables dual-domain disease profiling, where features are both globally and condition-specifically ranked to drive multi-risk classification accuracy. Once this step is complete, a labeled data frame is constructed where each instance (patient) is represented by a vector of refined features and a class label: 0 (No Risk) or 1 (Risk). This structured dataset is then forwarded to the third phase involving tri-layer classification using optimized learning models.

3.3. Tri-Step Classification Using Cm-Ro Optimized Hierarchical XGBoost for Risk Scoring

In the final phase of the Di-CVD Tri-Layer CX Classifier framework, the focus is on accurate, interpretable, and secure risk prediction using a novel Tri-Step Classification Strategy. This multi-layer pipeline integrates advanced feature optimization and decision-making strategies, culminating in the generation of individualized disease risk scores. The phase is composed of three tightly coupled steps: feature optimization, hierarchical classification, and prioritized risk scoring.

Step 1: Centroid Mutation-Rescue Optimization (Cm-Ro) Based Feature Selection

The first step begins by optimizing the feature set using a Centroid Mutation–Rescue Optimization (Cm-Ro) technique. Cm-Ro is inspired by population-based metaheuristics and aims to select feature subsets that best represent the discriminative nature of the disease classes. The technique operates by computing the centroid vector $\boldsymbol{C}$ of the feature matrix $X \in \mathbb{R}^{n \times m}$ in (where n is the number of patients and m is the number of features):

$$C = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{9}$$

Each feature vector is evaluated by its Euclidean distance from the centroid to measure its contribution to class separation. Mutation operations are then applied to introduce new combinations, followed by a rescue strategy that reintroduces high-impact features lost during mutation. This step ensures a multi-relational and multi-attribute configuration, meaning that cross-feature interactions (e.g., age × glucose, activity × BMI) are explicitly formed for more expressive 11odelling.

Step 2: Synergistic Hierarchical XGBoost Classification

The optimized feature set is then passed to a Synergistic Hierarchical XGBoost Classifier, which builds multiple tree-based learners in a layered decision process. The synergy here lies in the weighted prioritization of features like age, physical activity score, and disease duration (cycle years). These features are empirically found to have dominant predictive power and are assigned higher gain-based importance values in the model's objective function.

XGBoost's regularized loss function is given by:

$$\mathrm{L}(\phi) = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{k=1}^{K} \Omega(f_k) \tag{10}$$

where:

- $l\left(y_i, \hat{y}_i^{(t)}\right)$ is a differentiable convex loss function (e.g., logistic loss),
- $\hat{y}_i^{(t)}$ is the prediction of the ith patient at boosting round t,
- $\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \parallel w \parallel^2$ is the regularization term that penalizes tree complexity.

The hierarchy is achieved by training a sequence of classifiers, each refining its decisions based on the error residuals of the previous layer. Features are ranked and weighted dynamically using XGBoost's split-gain criterion:

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right] - \gamma \tag{11}$$

where:

- G and H are the first and second-order gradients of the loss function (w.r.t. predictions),
- L, R denote left and right splits of a node,
- λ *and* γ are regularization terms.

This ensures that the trees grow in a manner that reflects both discriminative strength and generalization, mitigating overfitting especially across IoT-derived noisy datasets.

Step 3: Synergistic Prioritized Risk Scoring

In the final step, the predicted outputs are post-processed to generate disease-specific risk scores using a Synergistic Prioritized Risk Scoring Algorithm. This step further deconstructs high-risk classifications into disease-wise sub-risks (Diabetes or CVD), using custom task-based weighting. For each patient, the final risk score R_s is computed using a linear priority-based fusion of features:

$$R_s = \sum_{j=1}^{m'} \omega_j . x_j \tag{12}$$

where:

- x_j is the value of the jthj^{th}jth selected feature,
- ω_j is the priority weight derived from domain-specific ranking (e.g., Di-CVD Rank),
- $m' \subset m$ is the reduced set of high-impact features after Cm-Ro and XGBoost.

Patients are thus not only classified as "At Risk" (1) or "No Risk" (0), but are also assigned an interpretable risk score that reflects the severity and source (DB or HD) of the health threat. This final output is transmitted securely to medical professionals, preserving data privacy while enhancing actionable insights for early intervention. Overall, the proposed model will be giving accurate individual risk scores in a secure manner without affecting the privacy of patients.

4. EXPERIMENTAL SETUP

4.1 Dataset Details

DB and HD-based attribute values presented dataset is given below.

https://catalog.data.gov/dataset/national-health-and-nutrition-examination-survey-nhanes

Interviews and physical examinations are also included in this dataset, which is NHANES. Every year, a nationally representative sample of roughly 5,000 people is examined by the survey.

These individuals are spread around the nation, with 15 of those counties receiving annual visits. Questions about diet, health, socioeconomic status, and demographics are all part of the NHANES interview. In addition to laboratory tests conducted by highly qualified medical professionals, the examination component includes medical, dental, and physiological assessments. CDC's Division for HD and Stroke Prevention (DHDSP) experts have calculated signs from this data source. The information has been stratified by age group, sex, and race/ethnicity and shown as trends.

4.2 HARDWARE REQUIREMENT

The following machine configuration will be used with the suggested Di-CVD Tri-Layer CX Classifier Based Risk Prediction Model in the MATLAB working platform (version R2019b):

Processor: Intel core i3

CPU Speed: 2.20 GHz

OS: Windows 7

RAM: 4GB

4.3 SOFTWARE REQUIREMENT

The MATLAB (matrix laboratory) working environment will be used to implement the suggested Di-CVD Tri-Layer CX Classifier Based RP Model. MATLAB is a fourth-generation programming language and multi-paradigm numerical computation environment. This was created especially for I/O and rapid and simple scientific computations. MathWorks created MATLAB, a proprietary programming language. User interface development, matrix manipulation, function and data visualisation, algorithm implementation, and interface with programs written in other languages, including C, C++, C#, Java, Fortran, and Python, are all made possible by MATLAB.

5.4 RESULTS AND DISCUSSION

This section presents a comprehensively experimental validation of the proposed Di-CVD Tri-Layer CX Classifier risk prediction model is provided using the NHANES database. The outcomes are investigated with different attributes embedded into the dataset. This new research has encouraging findings for the medical field, which raises hopes that these individuals may receive an early and effective diagnosis. In addition to classification accuracy (ACC), the classifier is evaluated using the average results for the classifiers and a few statistical metrics provided in the equation. SVM and XG BOOST are compared with the suggested Di-CVD Tri-Layer CX Classifier.

The ratio of accurately detected p”siti’e observations to all predicted positive observations is known as precision (P).

$$P = TP/TP+FP \tag{13}$$

The ratio of accurately identified positive observations to the total number of observations in the actual class is known as sensitivity or Recall (R).

$$R = TP/TP+FN \tag{14}$$

The weighted average of P and R is known as the F1 score. Consequently, FP and FN are required.

$$F1\ Score = 2*(R * P) / (R + P) \tag{15}$$

The following is how ACC is determined in terms of positives and negatives:

$$ACC = (TP+FP)/(TP+TN+FP+FN) \tag{16}$$

Where TP- True Positive

FP- False Positive

TN- True Negative

FN- False Negative

SVM, XGBOOST, and the suggested Di-CVSTLC are the 4 ML techniques that are compared in Table 1 utilising the assessment criteria of ACC, P, R and F-Measure.Among the three, the Support Vector Machine (SVM) approach shows the lowest performance, with an F-Measure of 81.47% and Accuracy of 83.54%, indicating limited effectiveness in balancing precision and recall. XGBOOST performs better, achieving 93.59% Precision, 87.25% Recall, and an improved F-Measure of 84.57%, demonstrating a more reliable classification performance than SVM. However, the proposed Di-CVSTLC model outperforms both existing methods across all metrics. Its greatest P (93.83%), R (89.68%), F-Measure (87.68%), and ACC (89.68%) show that it can accurately detect and classify cases more consistently. The Di-CVSTLC technique is a better option for the specified classification problem because of these results, which demonstrate its efficacy and resilience.

Table 1: Comparative Analysis of the suggested and current methods using various metrics

Methods	P (%)	R (%)	F-Measure (%)	ACC (%)
SVM	89.74	83.54	81.47	83.54
XGBOOST	93.59	87.25	84.57	87.25
Di-CVSTLC	93.83	89.68	87.68	92.18

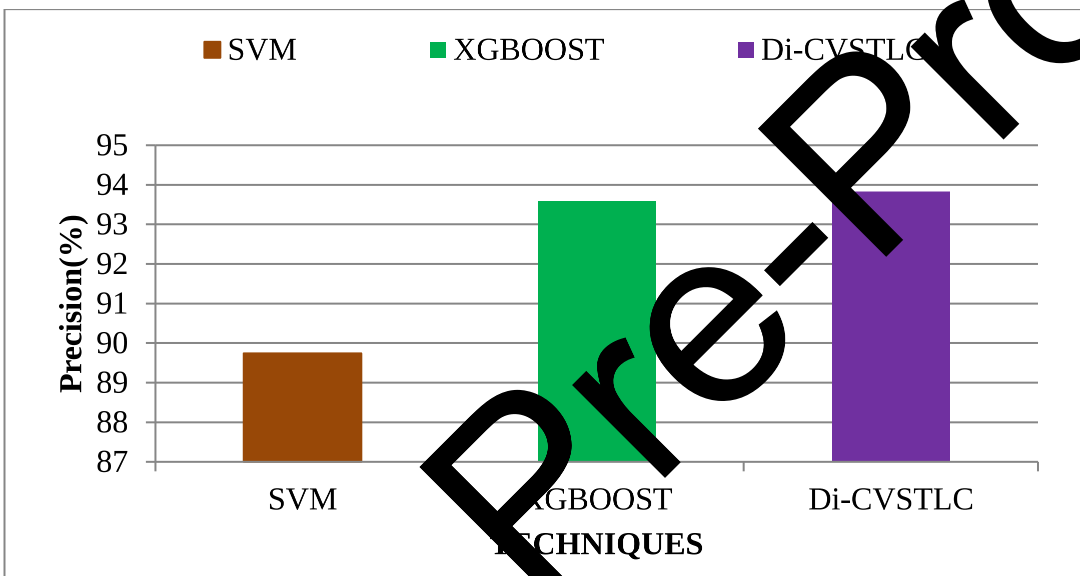


Figure 2: Comparison outcomes of the suggested DI-CVSTLC technique and current methods by P

Figure 2 shows the performance of the suggested DI-CVSTLC's p comparison results. The accuracy of SVM, XGBOOST, and the suggested Di-CVSTLC model in accurately identifying relevant (TP) cases out of all instances they predicted as positive is contrasted in the precision graph. In the graph, SVM shows a precision of 89.74%, indicating a relatively good ability to avoid false positives, but it still trails behind the other two models. XGBOOST performs better, with a precision of 93.59%, reflecting a higher level of accuracy in its positive predictions. However, the proposed Di-CVSTLC model achieves the highest precision at 93.83%, showing its superior capability to make highly accurate positive predictions. The small but noticeable improvement over XGBOOST suggests that Di-CVSTLC is more refined in its decision-making process, making fewer errors when identifying positive cases. The increasing precision values across the three models highlight the gradual improvement in predictive accuracy, with Di-CVSTLC emerging as the most reliable approach in terms of precision.

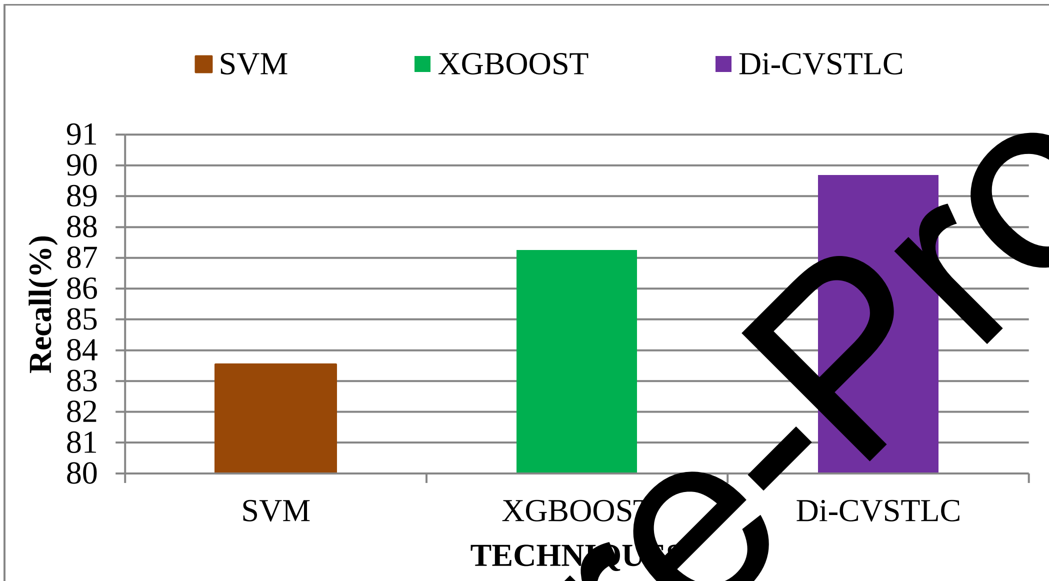


Figure 3: Comparison outcomes of the suggested DI-CVSTLC method and current approaches by R

Figure 3 shows the recall graph illustrates how effectively each model—SVM, XGBOOST, and the proposed Di-CVSTLC—identifies all relevant positive instances from the dataset. Among the three, SVM demonstrates the lowest recall at 83.54%, indicating that it misses a significant number of actual positive cases, leading to more false negatives. XGBOOST shows a noticeable improvement with a recall of 87.25%, suggesting a better ability to capture true positives. However, the proposed Di-CVSTLC model achieves the highest recall at 89.68%, outperforming both existing methods. This means Di-CVSTLC is the most effective at minimizing false negatives and ensuring that nearly all relevant instances are correctly identified.

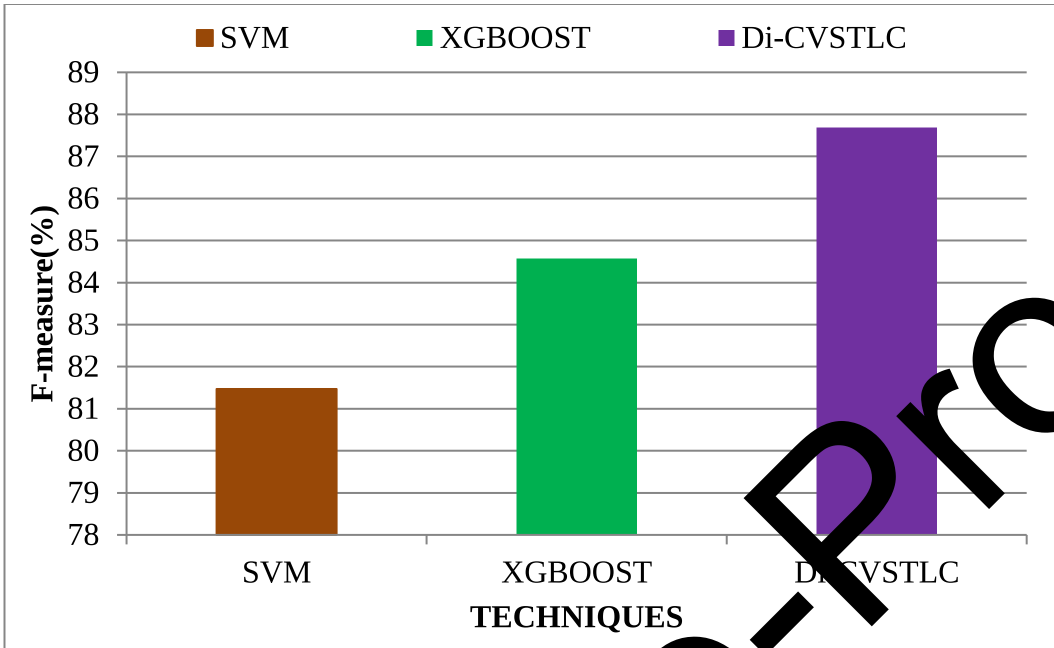


Figure 4: Comparison outcomes of the suggested Di-CVSTLC method and current approaches in F-measure

Figure 4 shows the F-Measure (or F1-Score) graph, which combines P and R into a single metric to provide an accurate representation of each model's performance. Among the three models, SVM records the lowest F-Measure at 81.47%, indicating weaker overall performance in maintaining a balance among identifying TP and avoiding FP or FN. XGBOOST shows a moderate improvement with an F-Measure of 84.57%, reflecting a better equilibrium between precision and recall. However, the proposed Di-CVSTLC model achieves the highest F-Measure at 87.68%, demonstrating its superior ability to maintain both high precision and high recall simultaneously. This high F-Measure score highlights Di-CVSTLC's consistency and robustness in classification tasks, making it more effective at delivering reliable results under varying data conditions. The increasing trend in the F-Measure values from SVM to XGBOOST to Di-CVSTLC clearly shows the progressive improvement in the overall quality of predictions, with Di-CVSTLC offering the best performance.

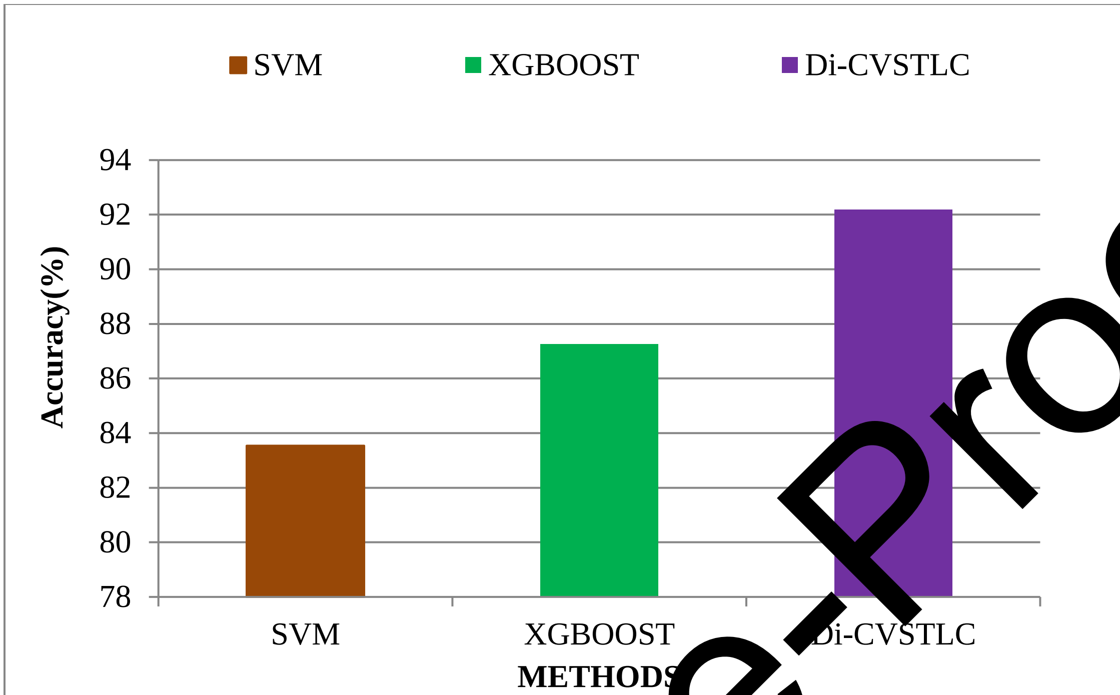


Figure 5: Comparison outcomes of the suggested Di-CVSTLC method and current approaches by ACC

Figure 5 shows the proposed DI-CVSTLC based classifiers give more accuracy than the existing classifier. SVM, XGBOOST, and the suggested Di-CVSTLC are the three models whose overall prediction ACC is compared in the ACC graph. SVM demonstrates the lowest accuracy at 83.54%, indicating that it correctly classifies a lower proportion of total instances compared to the other models. XGBOOST shows a significant improvement with an accuracy of 87.25%, reflecting its stronger ability to make correct predictions across both positive and negative classes. The proposed Di-CVSTLC model achieves the highest accuracy at 92.18%, clearly outperforming the existing approaches. This high accuracy score indicates that Di-CVSTLC consistently makes correct decisions and handles diverse data instances more effectively. The steady increase in accuracy from SVM to XGBOOST to Di-CVSTLC highlights the overall improvement in model performance, with Di-CVSTLC demonstrating the most reliable and accurate classification results among the three approaches.

6. CONCLUSION

In this study, a Di-CVD Tri-Layer CX Classifier-based framework was proposed for precise and secure risk prediction of diabetes and cardiovascular diseases in an IoT environment.

Unlike existing models, the proposed method integrates enhanced encryption techniques, intelligent feature ranking and physical activity factors, and a multi-phase classification strategy to ensure individual disease risk profiling. By leveraging the synergistic capabilities of the Cm-Ro optimized feature selection and hierarchical XGBoost model, the system improves both classification reliability and interoperability across heterogeneous IoT devices. Experimental evaluation on the NHANES dataset demonstrates high efficiency in terms of ACC (92.18%), R (89.68%), P (93.83%), and F1-score (87.68%) when compared with traditional classifiers like SVM and standard XGBoost. Future work will focus on expanding the model to multi-disease prediction systems with real-time deployment and federated learning-based privacy frameworks.

REFERENCES

[1] Balakumar, P., Maung-U, K., & Jagadeesh, G. (2016). Prevalence and prevention of cardiovascular disease and diabetes mellitus. Pharmacological research, 113, 600-609.

[2] Mamdiwar, S. D., Shakruwala, Z., Chadha, U., Srinivasan, K., & Chang, C. Y. (2021). Recent advances on IoT-assisted wearable sensor systems for healthcare monitoring. Biosensors, 11(10), 372.

[3] Abdulmalek, S., Nasir, A., Jabbar, W. A., Almuhaya, M. A., Bairagi, A. K., Khan, M. A. M., & Kee, S. H. (2022, October). IoT-based healthcare-monitoring system towards improving quality of life: A review. In Healthcare (Vol. 10, No. 10, p. 1993). MDPI.

[4] Ianculescu, M., Constantin, V. S., Gușatu, A. M., Petrache, M. C., Mihăescu, A. G., Bica, O., & Alexandru, A. (2025). Enhancing Connected Health Ecosystems Through IoT-Enabled Monitoring Technologies: A Case Study of the Monit4Healthy System. Sensors, 25([illegible]), 2292.

[5] Dash, S., Shakyawar, S. K., Sharma, M., & Kaushik, S. (2019). Big data in healthcare: management, analysis and future prospects. Journal of big data, 6(1), 1-25.

[6] Hong, L., Luo, M., Wang, R., Lu, P., Lu, W., & Lu, L. (2018). Big data in health care: Applications and challenges. Data and information management, 2(3), 175-197.

[7] Shehab, M., Abualigah, L., Shambour, Q., Abu-Hashem, M. A., Shambour, M. K. Y., Alsalibi, A. I., & Gandomi, A. H. (2022). Machine learning in medical applications: A review of state-of-the-art methods. Computers in Biology and Medicine, 145, 105458.

[8] Aldahiri, A., Alrashed, B., & Hussain, W. (2021). Trends in using IoT with machine learning in health prediction system. Forecasting, 3(1), 181-206.

[9] Ma, R., Gao, J., Mao, S. and Wang, Z., 2022. Association between heart rate and cardiovascular death in patients with coronary heart disease: A NHANES-based cohort study. Clinical Cardiology, 45(5), pp.574-582.

[10] Chen, L., Ming, J., Chen, T., Hébert, J.R., Sun, P., Zhang, L., Wang, H., Wu, Q., Zhang, C., Shivappa, N. and Ban, B., 2022. Association between dietary inflammatory index score and muscle mass and strength in older adults: a study from National Health and Nutrition Examination Survey (NHANES) 1999–2002. European Journal of Nutrition, pp.1-13.

[11] T. Thamaraimanalan, M. Mohankumar, S. Dhanasekaran, and H. Anandakumar, “Experimental analysis of intelligent vehicle monitoring system using Internet of Things (IoT),” EAI Endorsed Transactions on Energy Web, p. 169336, Jul. 2018, doi: 10.4108/eai.16-4-2021.169336.

[12] Zhang, S., Tian, J., Lei, M., Zhong, C. and Zhang, Y., 2022. Association between dietary fiber intake and atherosclerotic cardiovascular disease risk in adults: a cross-sectional study of 14,947 population based on the National Health and Nutrition Examination Surveys. BMC Public Health, 22(1), pp.1-9.

[13] Tu, Z.Z., Lu, Q., Zhang, Y.B., Shu, Z., Lai, Y.W., Ma, M.N., Xia, P.F., Geng, T.T., Chen, J.X., Li, Y. and Wu, L.J., 2022. Associations of Combined Healthy Lifestyle Factors with Risks of Diabetes, Cardiovascular Disease, Cancer, and Mortality Among Adults with Prediabetes: Four Prospective Cohort Studies in China, the United Kingdom, and the United States. Engineering.

[14] Zope, M., Binje, S., Johns, L. and Vasudevan, A., A System for Heart Disease Prediction Using Data Mining Techniques. International Journal of Innovations in Engineering Research and Technology, 3(4), pp.1-6.

[15] McKillop, A.L., 2017. Physical activity and exercise among patients with congenital heart disease: towards a model of pediatric cardiac rehabilitation (Doctoral dissertation, University of Toronto (Canada)).

[16] Ferdousi, R., Hossain, M.A. and El Saddik, A., 2021. Early-stage risk prediction of non-communicable disease using machine learning in health CPS. IEEE Access, 9, pp.96823-96837.