Indonesian Journal of Electrical Engineering and Computer Science
Vol. 34, No. 2, May 2024, pp. 1110~1121
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v34.i2.pp1110-1121

Empowering health data protection: machine learning-enabled diabetes classification in a secure cloud-based IoT framework

Dalia Ebrahim Hamid[1], Hanan M. Amer[2], Hossam El-Din Salah Moustafa[2], Hanaa Salem Marie[3]
[1]Faculty of Engineering, Delta University for Science and Technology, Gamasa, Egypt
[2]Department of Electronics and Communications, Faculty of Engineering, Mansoura University, Mansoura, Egypt
[3]Faculty of Artificial Intelligence, Delta University for Science and Technology, Gamasa, Egypt

Article Info

Article history:

Received Jan 2, 2024
Revised Feb 12, 2024
Accepted Feb 16, 2024

Keywords:

Classification
Cloud computing
Diabetes
Internet of things
Machine learning
Secure healthcare system

ABSTRACT

Smart medical devices and the internet of things (IoT) have enhanced healthcare systems by allowing remote monitoring of patient's health. Because of the unexpected increase in the number of diabetes patients, it is critical to regularly evaluate patients' health conditions before any significant illness occurs. As a result of transmitting a large volume of sensitive medical data, dealing with IoT data security issues remains a difficult challenge. This paper presents a secure remote diabetes monitoring (SR-DM) model that uses hybrid encryption, combining the advanced encryption standard and elliptic curve cryptography (AES-ECC), to ensure the patients' sensitive data is protected in IoT platforms based on the cloud. The health statuses of patients are determined in this model by predicting critical situations using machine learning (ML) algorithms for analyzing medical data sensed by smart health IoT devices. The results reveal that the AES-ECC approach has a significant influence on cloud-based IoT systems and the random forest (RF) classification method outperforms with a high accuracy of 91.4%. As a consequence of the outcomes obtained, the proposed model effectively establishes a secure and efficient system for remote health monitoring.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Hanaa Salem Marie
Faculty of Artificial Intelligence, Delta University for Science and Technology
Gamasa 35712, Egypt
Email: hana.salem@deltauniv.edu.eg

1. INTRODUCTION

Diabetes is a rapidly rising metabolic illness that is also one of the main causes of mortality globally. When pancreatic cells fail to generate sufficient insulin, blood sugar levels rise, wreaking havoc on most notably the eyes, a variety of organs, nerves, heart, and kidneys [1]. According to research conducted by Fitzmaurice *et al.* [2], the global prevalence of diabetes in 2017 was approximately 8.8%, and it is projected to increase to 9.9% by 2045. Furthermore, a recent study reveals the significant impact of diabetes, affecting approximately half a billion people worldwide, and estimates suggest that this number will rise by 25.0% to 51.0% between the years 2030 and 2045 [3].

Although there is no long-term treatment available for diabetes, it may be managed and controlled in case of early diagnosis. In such scenarios, the utilization of computer-aided technologies plays a crucial and beneficial role by facilitating precise medical decisions, thereby recommending timely and early essential treatments [4]. As a result, machine learning (ML) based developments make automated diabetes detection and diagnosis more likely and successful than the old way of manually diagnosing diabetes. Healthcare providers can use these predictions to tailor interventions, recommend lifestyle changes, and initiate early treatment strategies [5].

The advancement of the internet of things (IoT) and sensor technologies connected to medical wearable devices in recent years has improved patient care efficiency via intelligent and remote systems for health monitoring [6]. The integration of IoT with the cloud provides several resource management advantages, including powerful processing, resource distribution, facilitating user mobility in monitoring systems, and minimizing data fragmentation across multiple databases [7]. Current remote monitoring for healthcare systems in a cloud-based IoT environment comprises a setting in which biological data from patients is sent, stored, and shared to collect insights from anywhere and at any time [8]. While transferring medical data across IoT and storing it in the cloud, security and privacy issues have become critical concerns in these approaches. In general, health data are particularly sensitive to changes, and any changes in their contents might lead to mistakes in medical diagnosis [9].

Data security approaches such as cryptographic techniques are employed to encrypt data before its storage in the cloud, ensuring that even the cloud service provider is unable to access the data [10]. Cryptography could be used to ensure the integrity, confidentiality, and availability of data stored or accessed via the cloud. It converts plain data into encrypted or unreadable form for undesired users. Cryptographic methods involving encryption and decryption using keys. Two main encryption mechanisms are symmetric and asymmetric key encryption. In asymmetric encryption a public key is used for encryption and a private key is used for decryption, while in asymmetric encryption a single private key is used for both operations [11].

Hybrid cryptography refers to the combination of two or more cryptographic techniques to improve security that combines the strengths of both symmetric-key and asymmetric-key encryption schemes to address the limitations of each individual [12]. This study foresees the early stages of a diabetic patient using a strategy that does not need invasive procedures and uses ML approaches via a comprehensive secure remote diabetes monitoring (SR-DM) system that leverages both IoT and cloud technologies. This paper's primary contributions are as follows points:

- Propose an innovative hybrid cryptography approach that combines advanced encryption standard with elliptic curve cryptography (AES-ECC) to bolster the security of healthcare data during its transmission via cloud storage.
- Outline a methodology for disease prediction that employs various ML techniques and assesses the classification outcomes to detect diabetic mellitus at an early stage.
- The proposed approach results demonstrate that our proposed technique outperforms existing diabetes prediction systems in terms of both security and privacy.

2. RELATED WORK

This section reviews current literature on cloud-based IoT environments for health monitoring and prediction systems for identifying the health state of patients. Several papers profit from the advantages of combining IoT and cloud technology. Rahman *et al.* [13] developed a system for remote collection and analysis of patient physiological data. The smartphone and wearable sensors running the created application are used to gather, analyze, and upload data to the cloud server. Hosseinzadeh *et al.* [14] presented a predictive diagnosis framework for chronic kidney disease utilizing multimedia data collected through a cloud-based IoT platform. Deepika *et al.* [15] proposed a disease classification model by combining image processing with a secure cloud computing environment and an extended zigzag image encryption system that is resistant to various data threats.

Subashini *et al.* [16] implemented a smart system for the staging classification of cervical cancer images utilizing cloud platform and ML. It is created using real-time cervical images. Talib *et al.* [17] introduced smart records that are shared over the cloud. some sensors have been utilized to detect health data and then notify the person monitoring the patient's status via Telegram to the phone (e.g., oxygen percentage, heart rate, and body temperature). Kumar *et al.* [18] implemented an autonomic architecture for smart healthcare based on IoT and cloud platforms, which employed a random forest and logistic regression grid technique at edge nodes to analyze heart disease. Nigar *et al.* [19] introduced a hybrid strategy based on IoT and ML for monitoring and early detection of six major chronic diseases, including pneumonia, COVID-19, heart disease, diabetes, Alzheimer's, and brain tumor.

Stergiou *et al.* [20] suggested model employs four essential developing technologies: IoT, wireless sensor networks, cloud, and ML to identify harmful forms of viruses that infect humans or animals, generating global fear and disrupting human daily life. Paganelli *et al.* [21] presented an architecture for remote COVID-19 patient monitoring. The architecture takes into account data collection from users at home as well as hospital wards. Sahu *et al.* [22] presented a vital sign monitoring system based on a microcontroller unit linked to multiple sensors. It collects data from them and sends it to a smartphone application. It can analyze data and generate warnings based on sensor values. Security concerns were addressed in [15], [20] but not in others. In comparison to the analyzed studies, this paper introduces a secure approach in a cloud-based IoT platform using ML techniques for the early detection of diabetes, which employs a hybrid encryption approach and is a suitable solution for medical IoT resources, which has not been addressed in the studied publications.

3. PROPOSED MODEL

The suggested system's conceptual structure consists of the following phases. Initially, medical IoT devices receive biological data from patients. The medical datasets acquired are subsequently sent to the cloud subsystem via IoT and stored in a cloud database. To eliminate security issues, these medical data are encrypted before transmission. This encryption is done via a hybrid cryptography technique. Following encryption, the medical data will be sent to a cloud platform for disease diagnosis. The medical data diagnostic process employs suggested algorithms such as decision tree (DT), support vector machine (SVM), random forest (RF), k-nearest neighbors (KNN), and Naïve Bayes (NB). These algorithms classify the processed dataset into two categories: diabetic and non-diabetic. The cloud classification results are sent to doctors or specialists in the context of a person's health using IoT. Figure 1 depicts the suggested SR-DM paradigm in a cloud-based IoT context that takes advantage of a hybrid encryption mechanism. The combination of all of these problems is examined in this study when it comes to the progression of diabetic disease. As a result, the primary goal of this study is to provide a safe health monitoring model for early detection of diabetes based on anticipating the crucial patient's status.

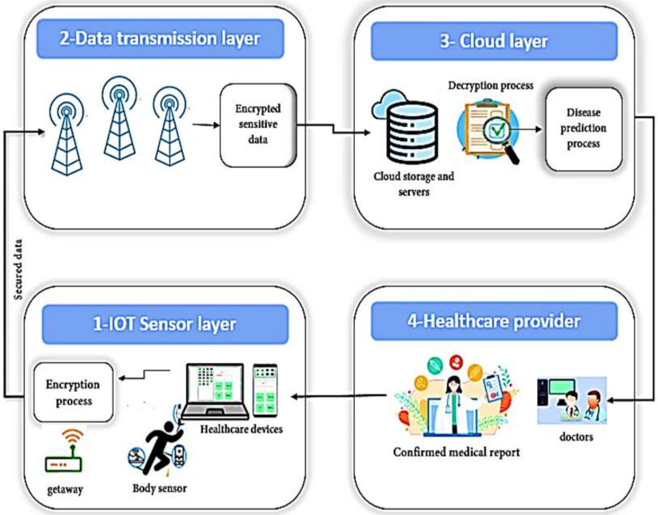


Figure 1. The proposed (SR-DM) model in a cloud-based IoT platform

Step 1: the IoT sensor layer integrated into wearable devices gathers information from various medical sensors, location sensors, body sensors, and ambient sensors. Depending on the nature of the problem statement, wearable devices may include sensors for insulin levels, glucose, body temperature, electrocardiograms (ECG), blood pressure, blood oxygen levels, pulse rate, galvanic skin response, and electromyography. Data from these sensors, along with modern data analytics and ML algorithms, can improve the accuracy and efficiency of diabetes prediction in IoT devices. Because health IoT sensor network devices are more sensitive to security threats than other network devices, an element is built to provide security requirements for IoT data. Before uploading the acquired sensitive data to the cloud, the obtained IoT data is encrypted using a hybrid manner. Step 2: the medical data of the patients is delivered to the cloud storage for diagnosis of the patient's health condition through the data transmission layer. This section must supply secret shares to transmit them to cloud servers as part of a distributed data storage framework. Step 3: whereas the cloud stores the medical data forwarded by medical IoT sensors to patients. This section also works with supplying and providing services to the associated users, which include healthcare professionals and doctors. These services can be integrated with predicting potential diseases using ML algorithms. Step 4: healthcare providers include physicians, hospitals, and emergency responders. Doctors can utilize the transmitted diagnosis results to review and validate them before making any medical suggestions to patients.

4. METHOD

The operations outlined above are executed within our proposed model through the procedural flow depicted in Figure 1, the collection of sensor data from IoT medical devices and IoT device data occurs simultaneously. Subsequently, a hybrid encryption method is applied to the gathered medical data, followed by the secure transfer of medical data to communication service providers. The data is then transferred and stored in cloud repositories as distributed data storage. Sequentially, the process involves decrypting secured data, performing data preprocessing, and ultimately predicting the occurrence of disease.

4.1. Providing the security of data in the proposed model

Due to security being a paramount concern in systems established in the IoT environment, sensitive patients' medical data is encrypted using performing in order to provide patient anonymity, confidentiality, and security needs. AES-ECC hybrid encryption combines the symmetric encryption strength of AES with the asymmetric encryption efficiency of ECC. It combines the speed and efficiency of ECC's small key size with the robust encryption capabilities of AES, resulting in improved performance and reduced key size for data protection in the cloud. ECC employs key standards to secure data with a smaller key size. When used alongside AES, it offers effective data protection. ECC determines the key size for generating ciphertext, which is then utilized by AES. The use of ECC for key exchange adds an extra layer of security and efficiency to the encryption process. The proposed AES-ECC provides secure health data sharing through the cloud environment by providing confidentiality, privacy, and integrity.

The hybrid proposed approach depicted in Figure 2 provides a robust solution for securing IoT-to-cloud communications by combining AES with ECC. The graphical representation highlights the proposed technique, showcasing secure data transfer from IoT devices to the cloud server, followed by secure storage of encrypted data. Moreover, the level of innovation can be evaluated based on factors such as computational time and cost. The proposed approach also demonstrates strong security measures in preventing attacks. For instance, if an attacker tries to access personal information from the user's side, the input file undergoes AES encryption, resulting in complete encryption. As a result, even if the attacker successfully acquires the user-uploaded file, it becomes futile since the information was already encrypted during the upload process. Likewise, in the event of an attack, the encrypted file remains incomprehensible to the attacker, thereby ensuring the security of the data and protecting it against unauthorized access attempts.

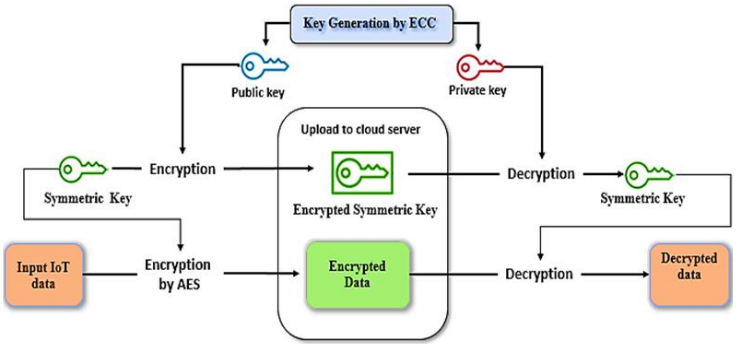


Figure 2. Representation of AES-ECC algorithm

The procedures involved in encrypting data from IoT to cloud server using the suggested technique AES-ECC hybrid cryptography are as follows:

- Generate asymmetric key pair (ECC). Start by generating an ECC key pair for each interacting entity, specifically the IoT device and cloud server. This pair consists of two keys: one public and one private. The public key is publicly shared, whilst the private key is safely stored by the corresponding entity.
- IoT device initiates communication. When an IoT device wants to communicate with the cloud, it initiates the process by requesting the cloud server's ECC public key. This request is made through a secure channel to ensure the authenticity of the received public key.
- Cloud server shares ECC public key. Cloud server responds to IoT device's request by sharing its ECC public key. IoT device receives this public key and uses it for the next steps in the encryption process.
- Generate symmetric key (AES). IoT device generates a random symmetric key for use with AES encryption. This key is known as the session key and will be used exclusively for the current communication session.

- Encrypt data with AES. The actual data that needs to be transmitted from the IoT device to the cloud is encrypted using the AES algorithm and the symmetric session key generated in the previous step. AES is a symmetric encryption algorithm known for its efficiency and strength.
- Encrypt the symmetric key with ECC public key. The IoT device then encrypts the symmetric session key using the ECC public key obtained from the cloud server. This step is performed using ECC asymmetric encryption, ensuring that only the cloud server, with its corresponding private key, can decrypt the session key.
- Transmit encrypted symmetric key. IoT device transmits the encrypted symmetric session key to a cloud server. Even if intercepted during transmission, the encrypted symmetric key remains secure because only a cloud server can decrypt it using its private key.
- Cloud server decrypts symmetric key. Upon receiving the encrypted symmetric session key, the cloud server uses its private key (ECC asymmetric decryption) to decrypt and obtain the symmetric session key. Now both IoT device and cloud server possess the same session key for symmetric encryption and decryption.
- Decrypt data with AES. The cloud server uses the decrypted symmetric session key to decrypt and retrieve the original data.

4.1.1. Algorithm for the proposed AES-ECC

a. Generation of a public key using ECC

```
Select an initial number (x).
Choose another number (x(i)) to generate the public key, where x(i) is less than x.
Determine a point on the curve, denoted as Q, where Q is greater than x.
Calculate the public key (P) by multiplying x(i) with Q: P = x(i) × Q.
After calculation, return the public key P.
```

b. Data encryption and decryption by AES

```
taking the input medical data.
Add the public key produced by ECC to the input data.
Encrypt the input data using AES encryption and the ECC-generated public key.
Upload the encrypted data to the server after AES encryption.
Upon downloading the data from the server, use the ECC-provided public key to decrypt and
restore the original data.
forward the decrypted medical data for storing in the cloud for authorized access.
```

4.2. The architecture of the proposed healthcare system's medical data classification module

Based on their vital signs, ML methods are used to categorize patients as healthy or sick. The cloud layer delivers the patient a diagnostic and emergency warning message whenever the patient's health condition is diagnosed during classification. The proposed framework to develop the ML model for diabetes prediction is shown in Figure 3. It explains the method used to train or update the training model and predict the incidence/prevalence of diabetes. The following explains the workflow of the training/update training and prediction in detail.

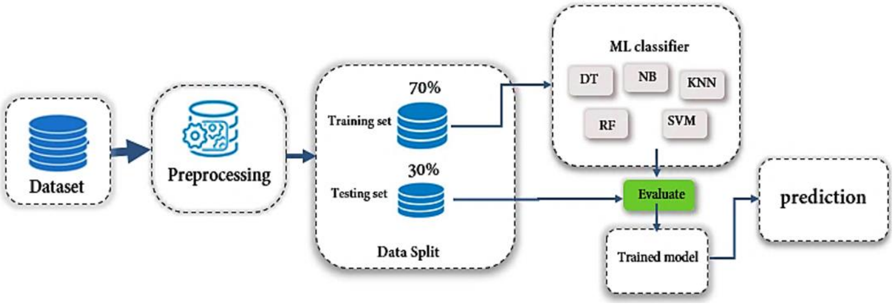


Figure 3. Framework of proposed health condition monitoring

4.2.1. Dataset pre-processing

The proposed algorithm's pre-processing stage includes outlier rejection (P), missing value imputing (Q), and feature selection. An outlier refers to an observation that deviates from the rest of the data points in a dataset [23]. Classification algorithms are impacted by the distribution of data; hence, data that deviates from the distribution must be excluded. To eliminate outliers, we employed the interquartile range (IQR) approach. The mathematical formula for outlier rejection is presented in (1).

$$p(x) = \begin{cases} x \; if \; Q1 - 1.5 * IQR \leq x \leq Q3 + 1.5 * IQR \\ \quad reject \; otherwise \end{cases} \tag{1}$$

Where x signifies the dimension space M enrolled feature vector (FV) instances, such as x ∈ R^m. Knowing that Q1, Q3, and the IQR belong to R^m, Q1 denotes the initial quartile, Q3 indicates the third quartile, and the IQR is the band of interquartile of the utilized qualities.

After removing outliers, the attributes were further processed to fill in any missing or null values [24]. The presence of missing or null values can impact the accuracy of predictions made by any classifier. In the proposed framework, instead of dropping the instances with missing values, imputation was performed using the mean values of the attributes, as indicated in (2). Replacement by mean is advantageous since it allows for the imputation of continuous data without adding outliers.

$$Q(a) = \begin{cases} mean(a) \;, if \; a = null/missed \\ \quad a \;, otherwise \end{cases} \tag{2}$$

Where a is the number of occurrences of the feature vector in n-dimensional space, a ∈ R^m.

Feature selection offers the advantage of improving correlation. As the dimension of the features increases, the accuracy of classifiers tends to improve as well. However, when the feature domain expands without a corresponding increase in the sample size, it becomes necessary to enhance the effectiveness of the classifiers to maintain their performance. Because of the curse of dimensionality, the feature space grows increasingly crowded, pushing classifiers to become overfitted and lose functional generalizability [25]. This literature used methods for the feature selection, namely the correlation-based technique.

4.2.2. Machine learning classification models

The classification process involves training the model on historical datasets containing labeled information about individuals with and without diabetes. The model learns patterns and trends from this data, allowing it to generalize and make predictions on new, unseen data [26]. The proposed framework incorporates the training and testing of multiple ML models, including KNN, NB, DT, RF, and SVM which are employed for their interpretability and effectiveness in capturing relationships between various risk factors.

i) KNN classifies an individual based on the majority class among its k-nearest neighbors in the feature space. KNN can be applied to diabetes diagnosis by considering the similarity of an individual's health features to those of its nearest neighbors with known diabetes status.

ii) NB is a probabilistic classifier that calculates the likelihood of a certain class given the observed features, assuming feature independence. NB can be used to estimate the probability of diabetes based on the observed health features, making it suitable for medical diagnosis.

iii) DT is a tree-like model where each node represents a decision based on a feature. DTs can be used to create a decision-making structure for diabetes diagnosis, splitting the data based on relevant health features.

iv) RF is an ensemble learning method that builds multiple decision trees and combines their predictions. RF can be employed for diabetes diagnosis by capturing complex relationships in health data and providing robust predictions.

v) SVM constructs a hyperplane to maximize the margin between different classes in the feature space. SVM can be applied to diabetes diagnosis by finding an optimal hyperplane that separates individuals with and without diabetes, considering various health features.

5. RESULTS AND DISCUSSION

The essential goal of this study involves accurately categorizing the data based on its health status, effectively distinguishing between healthy samples and those indicating the presence of a disease, and considering the severity of the disease. Additionally, ensuring secure data transfer is of utmost importance. The implementation of this system is carried out using the Python programming language. The experiments conducted utilized the pima indian diabetes dataset (PIDD) to evaluate the proposed work based on several parameters, namely accuracy (Acc), sensitivity (Sn), precision (P), and area under the receiver operating characteristic curve (AUC-ROC). Furthermore, the security of the proposed system is assessed by analyzing encryption time and decryption time. Initially, a comparison is made between the security of the proposed system and the prevailing methodology. Various experiments employing diverse classification algorithms are conducted to classify instances of disease. The resulting experimental outcomes are also tested to evaluate the performance of the proposed technique.

5.1. Dataset

The PIDD was used as the primary dataset in this study, sourced from the University of California, Irvine (UCI) machine repository standard dataset. The PIDD consists of records from 768 female diabetic patients, encompassing 8 distinct features are pregnancies (C1), glucose (C2), blood pressure (C3), skin thickness (C4), insulin (C5), BMI (C6), diabetes pedigree function (C7), and age (C8)]. The dataset is divided into two groups for analysis: 500 normal individuals without diabetes (0) and is labeled (1) for 268 diabetic patients. Figure 4 shows the population distribution of all features in the PIDD where blue denotes non-diabetes, and orange color denotes diabetes classification.

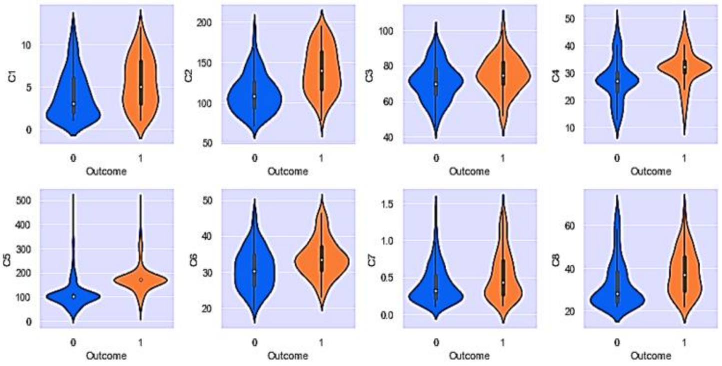
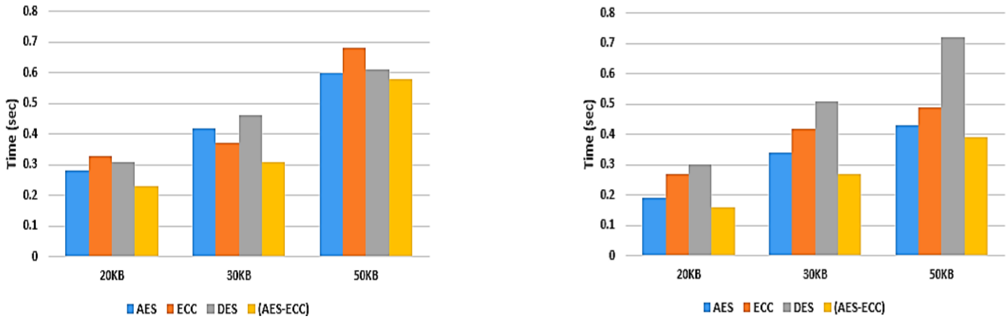


Figure 4. The population distribution of all features in the PIDD

5.2. Evaluation of the performance of data security for the proposed system

Encryption time refers to the duration it takes for the encryption process to complete, measured from the start to the end of the process. It represents the time required for the encryption algorithm to convert plaintext data into ciphertext. Decryption time is calculated by subtracting the start time from the end time of the decryption process. It indicates the time required to decrypt the ciphertext and recover the original plaintext.

Figure 5 illustrates a graphical representation of the performance of the proposed AES-ECC algorithm compared to conventional encryption algorithms AES, ECC, and DES in terms of encryption time and decryption time. In the analysis, files ranging from 20 KB to 50 KB were evaluated. Figure 5(a) demonstrates the encryption time comparison between AES, ECC, and DES approaches. For a 20 KB file, the proposed AES-ECC took 0.23 seconds to encrypt, while the existing AES, ECC, and DES algorithms took 0.28 seconds, 0.33 seconds, and 0.31 seconds respectively for the same encryption task. Moreover, the proposed approach demonstrates superior performance for file sizes of 30 KB and 50 KB as well. Figure 5 (b), it can be observed that the proposed technique achieves a decryption time of 0.16 seconds for a file size of 20 KB. In comparison, the existing methods demonstrate lower performance, requiring more time for decryption. For a 30 KB file size, the proposed technique demonstrates a decryption time of 0.27 seconds, which is faster than the current methodology. The overall analysis suggests that the proposed technique outperforms existing methods in terms of performance.



(a) (b)

Figure 5. Performance of the proposed AES-ECC (a) encryption time analysis (b) decryption time analysis

5.2.1. Comparisons of proposed hybrid approach AES-ECC with various algorithms

Different algorithms are examined for functionality and space optimization in IoT-cloud platforms. Table 1 shows a comparison of several algorithms based on various parameters. Cryptographic algorithms have been compared for performance assessment based on the number of keys used, keys in bits, rounds, block size, and security. Hybrid encryption AES-ECC offers the advantage of merging the strengths of both symmetric and asymmetric encryptions. It combines the swiftness and effectiveness of symmetric encryption with the security and adaptability of asymmetric encryption. This approach excels at managing extensive data while maintaining security and performance. Additionally, it solves the key distribution challenge of symmetric encryption by securely exchanging the symmetric key through asymmetric encryption.

Table 1. Cryptography algorithm comparison

	DES	AES	RSA	ECC	Blowfish	AES-ECC
No. of key	1	1	2	2	1	1
Key length	56 bits	128,192,256 bits	1,024 bits	160 bits	32 to 448 bits	64 to 256 bits
Cipher type	Symmetric	Symmetric	Asymmetric	Asymmetric	Symmetric	Symmetric and asymmetric
Rounds	16	key:128 bits -10 key:192 bits -12 key:256 bits -14	1	16	16	10
Block size	64 bits	128 bits	Min 512 bits	64 bits	64 bits	128 bits
Security	Not secure enough	Adequately secured	Least secure	Adequately secured	Least secure	High secure

5.3. Performance analysis of classification module

Different performance evaluation metrics were employed to evaluate classifier performance [27]. The confusion matrix is used to calculate these measures. The binary classification matrix is shown Figure 6. We calculated the following performance evaluation metrics from Figure 6 and illustrated them mathematically in (4) to (6).

- Accuracy (Acc): measures the model's percentage of correct predictions.

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

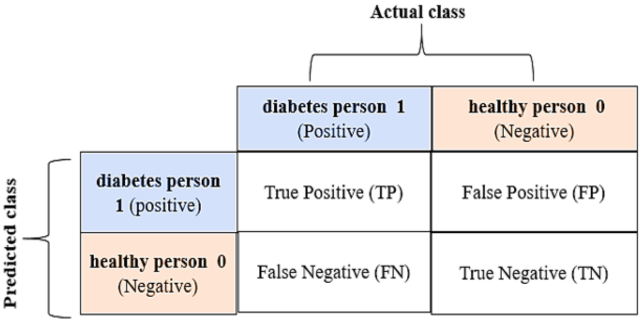
- Precision (P): measures the proportion of true positives among the instances predicted as positive by the model.

$$= \frac{TN}{TN+FP} \tag{5}$$

- Sensitivity (Sn): measures the proportion of true positives among the positive instances.

$$Sn = \frac{TP}{TP+FN} \tag{6}$$

- AUC-ROC: a graphical representation of the model's ability to distinguish between positive and negative instances across different thresholds. This assessment metric is commonly used to evaluate binary classification models. It measures the area under the ROC curve, which depicts the relationship between the true positive rate (TPR) and the false positive rate (FPR) at different classification thresholds.

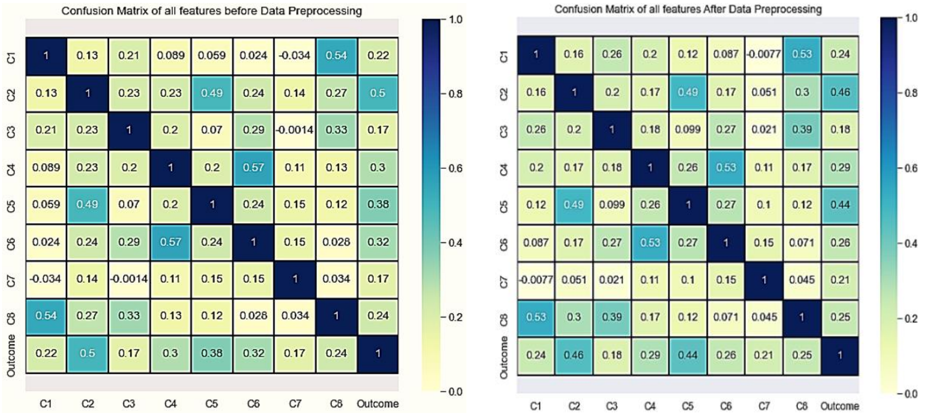


	diabetes person 1 (Positive)	healthy person 0 (Negative)
diabetes person 1 (positive)	True Positive (TP)	False Positive (FP)
healthy person 0 (Negative)	False Negative (FN)	True Negative (TN)

Figure 6. Confusion matrix

5.3.1. Confusion matrix

The correlation-based confusion matrix as shown in Figure 7 represents the outcomes of outlier rejection and missing value imputation. Both qualitative and quantitative analyses of Figure 7(a) and Figure 7(b) indicate an enhancement in the correlation between attributes and the target outcome after the application of outlier rejection and missing value filling. Notably, the correlation coefficients for attributes C3, C4, and C5 exhibit significant improvement. This improved correlation is advantageous for correlation-based feature selection techniques.



Confusion Matrix of all features before Data Preprocessing

	C1	C2	C3	C4	C5	C6	C7	C8	Outcome
C1	1	0.13	0.21	0.089	0.059	0.024	-0.034	0.54	0.22
C2	0.13	1	0.23	0.23	0.49	0.24	0.14	0.27	0.5
C3	0.21	0.23	1	0.2	0.07	0.29	-0.0014	0.33	0.17
C4	0.089	0.23	0.2	1	0.2	0.57	0.11	0.13	0.3
C5	0.059	0.49	0.07	0.2	1	0.24	0.15	0.12	0.38
C6	0.024	0.24	0.29	0.57	0.24	1	0.15	0.028	0.32
C7	-0.034	0.14	-0.0014	0.11	0.15	0.15	1	0.034	0.17
C8	0.54	0.27	0.33	0.13	0.12	0.028	0.034	1	0.24
Outcome	0.22	0.5	0.17	0.3	0.38	0.32	0.17	0.24	1

(a)

Confusion Matrix of all features After Data Preprocessing

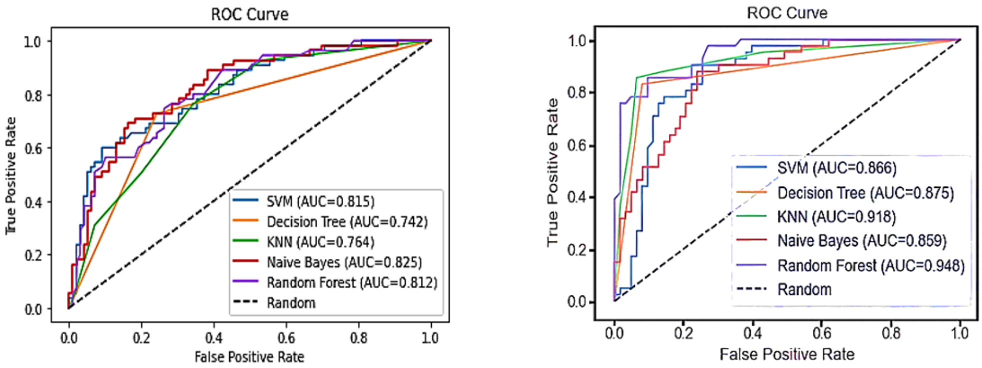
	C1	C2	C3	C4	C5	C6	C7	C8	Outcome
C1	1	0.16	0.26	0.2	0.12	0.087	-0.0077	0.53	0.24
C2	0.16	1	0.2	0.17	0.49	0.17	0.051	0.3	0.46
C3	0.26	0.2	1	0.18	0.099	0.27	0.021	0.39	0.18
C4	0.2	0.17	0.18	1	0.26	0.53	0.11	0.17	0.29
C5	0.12	0.49	0.099	0.26	1	0.27	0.1	0.12	0.44
C6	0.087	0.17	0.27	0.53	0.27	1	0.15	0.071	0.26
C7	-0.0077	0.051	0.021	0.11	0.1	0.15	1	0.045	0.21
C8	0.53	0.3	0.39	0.17	0.12	0.071	0.045	1	0.25
Outcome	0.24	0.46	0.18	0.29	0.44	0.26	0.21	0.25	1

(b)

Figure 7. The generated confusion matrix depicts the PIDD-based feature correlation (a) before data preprocessing and (b) after data preprocessing

5.3.2. ROC curve

The performance variation of the models can be observed in Figure 8, specifically focusing on the AUC metric. Figure 8(a) represents the models' performance without any pre-processing, while Figure 8(b) illustrates the models' performance with suitable pre-processing techniques applied. It is evident from Figure 8(b) that the RF model outperforms the DT, SVM, NB, and KNN models. The AUC values for the RF classifier on the test set demonstrate its robustness and effectiveness in utilizing this dataset.



(a) (b)

Figure 8. ROC curve of the classifiers (a) without data preprocessing and (b) proposed model

5.3.3. Comparison of the proposed model for diabetes diagnosis with other works

Various evaluation criteria are taken into account to analyze ensemble learners and ML models. This section of the research involves a comparison of the results obtained by previous authors in the context of diabetes classification. The study specifically compares the performance of ML models using the PIDD. The findings suggest that RF classifier for the proposed model achieves the highest Acc, P, Sn, and AUC values, reaching 91.40%, 89.70%, 85.40%, and 94.80%, respectively, surpassing other ML algorithms. Additionally, Table 2 provides a comparative analysis of the results obtained by previous researchers.

Table 2. The performance of the proposed model is compared to recent approaches

Author	Classifier	Acc (%)	P (%)	Sn (%)	AUC (%)
Bhoi *et al.* [28]	DT	70.80	70.10	70.82	64.80
	SVM	66.50	67.10	66.50	70.70
	KNN	71.11	70.30	71.10	73.70
	NB	73.60	74.50	73.60	81.80
	RF	75.40	75.10	75.40	80.80
	LR	76.80	73.30	76.80	82.50
Ramesh *et al.* [29]	KNN	79.8	-	87.20	-
	LR	73.30	-	70.20	-
	NB	73.10	-	66.60	-
Perdana *et al.* [30]	SVM+RBF	83.20	-	87.30	-
	KNN	83.12	-	74.19	-
The proposed model	SVM	82.70	78.00	78.00	86.60
	DT	88.50	87.20	82.90	87.50
	KNN	86.50	84.60	80.50	91.80
	NB	75.00	72.70	58.50	85.90
	RF	91.40	89.70	85.40	94.80

6. CONCLUSION

Regarding the huge increase in the number of diabetics, the growing requirement for cloud-based IoT platforms for healthcare monitoring and disease prediction systems, maintaining patient privacy and ensuring the security of sensitive healthcare data have become significant challenges. To address these concerns, this study proposes an effective strategy that prioritizes patient privacy while leveraging healthcare data for disease prediction within the current healthcare system. The proposed system (SR-DM) incorporates a novel approach by integrating AES and ECC techniques. To assess the system's performance, both secure data transmission and classification performance analyses were conducted. The proposed AES-ECC algorithm's performance was evaluated by comparing it with conventional encryption algorithms like DES, AES, and ECC. The evaluation considered factors such as encryption time and decryption time analysis. Additionally, the performance of the RF classifier was assessed in comparison to existing classifiers such as NB, DT, KNN, and SVM. The RF classifier achieved an impressive performance of the other classifiers with 91.40% accuracy, 89.70% precision, 85.40% sensitivity, and 94.80% AUC. Experimental results confirm that the proposed approach outperforms existing systems in terms of disease prediction accuracy while also providing enhanced privacy and security measures. In future work, the model can be further improved by incorporating more generalized strategies to handle diverse dataset types beyond those collected by the IoT. Deep learning techniques can also be explored to enhance disease prediction outcomes while maintaining robust security and privacy measures.

REFERENCES

[1] N. A. ElSayed *et al.*, “2. classification and diagnosis of diabetes: standards of care in diabetes—2023,” *Diabetes Care*, vol. 46, no. Supplement_1, pp. S19–S40, Jan. 2023, doi: 10.2337/dc23-S002.

[2] C. Fitzmaurice *et al.*, “Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 32 cancer groups, 1990 to 2015: a systematic analysis for the global burden of disease study,” *JAMA Oncology*, vol. 3, no. 4, p. 524, Apr. 2017, doi: 10.1001/jamaoncol.2016.5688.

[3] P. Saeedi *et al.*, “Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: results from the International Diabetes Federation Diabetes Atlas, 9th edition,” *Diabetes Research and Clinical Practice*, vol. 157, p. 107843, Nov. 2019, doi: 10.1016/j.diabres.2019.107843.

[4] A. Hennebelle, H. Materwala, and L. Ismail, “HealthEdge: a machine learning-based smart healthcare framework for prediction of type 2 diabetes in an integrated IoT, edge, and cloud computing system,” *Procedia Computer Science*, vol. 220, pp. 331–338, 2023, doi: 10.1016/j.procs.2023.03.043.

[5] S. Dasari, B. Poonguzhali, and M. S. Rayudu, “An efficient machine learning approach for classification of diabetic retinopathy stages,” *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 30, no. 1, p. 81, Apr. 2023, doi: 10.11591/ijeecs.v30.i1.pp81-88.

[6] S. D. Mamdiwar, A. R, Z. Shakruwala, U. Chadha, K. Srinivasan, and C.-Y. Chang, “Recent advances on IoT-assisted wearable sensor systems for healthcare monitoring,” *Biosensors*, vol. 11, no. 10, p. 372, Oct. 2021, doi: 10.3390/bios11100372.

[7] R. Kumar and N. Agrawal, “Analysis of multi-dimensional Industrial IoT (IIoT) data in edge–fog–cloud based architectural frameworks : a survey on current state and research challenges,” *Journal of Industrial Information Integration*, vol. 35, p. 100504, Oct. 2023, doi: 10.1016/j.jii.2023.100504.

[8] S. El Kafhali and I. El Mir, “Exploring the effectiveness of cloud, internet of things and fog computing for healthcare monitoring systems,” in *Computational Intelligence for Medical Internet of Things (MIoT) Applications*, Elsevier, 2023, pp. 77–91. doi: 10.1016/B978-0-323-99421-7.00008-8.

[9] H. N. Saha and S. Debnath, “Security and privacy of IoT devices in healthcare systems,” in *Smart Healthcare System Design*, Wiley, 2022, pp. 143–165. doi: 10.1002/9781119792253.ch7.

[10] A. S. Babrahem and M. M. Monowar, “Preserving confidentiality and privacy of the patient’s EHR using the OrBAC and AES in cloud environment,” *International Journal of Computers and Applications*, vol. 43, no. 1, pp. 50–61, Jan. 2021, doi: 10.1080/1206212X.2018.1505025.

[11] V. Agarwal, A. K. Kaushal, and L. Chouhan, “A survey on cloud computing security issues and cryptographic techniques,” 2020, pp. 119–134. doi: 10.1007/978-981-15-2071-6_10.

[12] S. Ahmad, S. Mehfuz, and J. Beg, “Hybrid cryptographic approach to enhance the mode of key management system in cloud environment,” *The Journal of Supercomputing*, vol. 79, no. 7, pp. 7377–7413, May 2023, doi: 10.1007/s11227-022-04964-9.

[13] M. J. Rahman, B. I. Morshed, B. Harmon, and M. Rahman, “A pilot study towards a smart-health framework to collect and analyze biomarkers with low-cost and flexible wearables,” *Smart Health*, vol. 23, p. 100249, Mar. 2022, doi: 10.1016/j.smhl.2021.100249.

[14] M. Hosseinzadeh *et al.*, “A diagnostic prediction model for chronic kidney disease in internet of things platform,” *Multimedia Tools and Applications*, vol. 80, no. 11, pp. 16933–16950, May 2021, doi: 10.1007/s11042-020-09049-4.

[15] J. Deepika, C. Rajan, and T. Senthil, “Security and privacy of cloud-and IoT-based medical image diagnosis using fuzzy convolutional neural network,” *Computational Intelligence and Neuroscience*, vol. 2021, pp. 1–17, Mar. 2021, doi: 10.1155/2021/6615411.

[16] P. Subashini, T. T. Dhivyaprabha, M. Krishnaveni, and M. B. Jennyfer Susan, “Smart intelligent system for cervix cancer image classification using google cloud platform,” in *Enabling Technologies for Effective Planning and Management in Sustainable Smart Cities*, Cham: Springer International Publishing, 2023, pp. 245–281. doi: 10.1007/978-3-031-22922-0_10.

[17] S. H. Talib, L. A. Abdul-Rahaim, A. J. Alrubaie, and I. M. Raseed, “Design smart hospital system based on cloud computing,” *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 29, no. 2, p. 797, Feb. 2023, doi: 10.11591/ijeecs.v29.i2.pp797-807.

[18] M. Kumar, A. Rai, Surbhit, and N. Kumar, “Autonomic edge cloud assisted framework for heart disease prediction using RF-LRG algorithm,” *Multimedia Tools and Applications*, vol. 83, no. 2, pp. 5929–5953, Jan. 2024, doi: 10.1007/s11042-023-15736-9.

[19] N. Nigar, A. Jaleel, S. Islam, M. K. Shahzad, and E. A. Affum, “IoMT meets machine learning: from edge to cloud chronic diseases diagnosis system,” *Journal of Healthcare Engineering*, vol. 2023, pp. 1–13, Jun. 2023, doi: 10.1155/2023/9995292.

[20] C. L. Stergiou, A. P. Plageras, V. A. Memos, M. P. Koidou, and K. E. Psannis, “Secure monitoring system for IoT healthcare data in the cloud,” *Applied Sciences*, vol. 14, no. 1, p. 120, Dec. 2023, doi: 10.3390/app14010120.

[21] A. I. Paganelli *et al.*, “A conceptual IoT-based early-warning architecture for remote monitoring of COVID-19 patients in wards and at home,” *Internet of Things*, vol. 18, p. 100399, May 2022, doi: 10.1016/j.iot.2021.100399.

[22] M. L. Sahu, M. Atulkar, M. K. Ahirwal, and A. Ahamad, “Vital sign monitoring system for healthcare through IoT based personal service application,” *Wireless Personal Communications*, vol. 122, no. 1, pp. 129–156, Jan. 2022, doi: 10.1007/s11277-021-08892-4.

[23] A. Boukerche, L. Zheng, and O. Alfandi, “Outlier detection: methods, models, and classification,” *ACM Computing Surveys*, vol. 53, no. 3, pp. 1–37, May 2021, doi: 10.1145/3381028.

[24] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, “A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data,” *Frontiers in Energy Research*, vol. 9, Mar. 2021, doi: 10.3389/fenrg.2021.652801.

[25] P. Dhal and C. Azad, “A comprehensive survey on feature selection in the various fields of machine learning,” *Applied Intelligence*, vol. 52, no. 4, pp. 4543–4581, Mar. 2022, doi: 10.1007/s10489-021-02550-9.

[26] I. Arief Wisky, M. Yanto, Y. Wiyandra, H. Syahputra, and F. Hadi, “Machine learning classification of infectious disease distribution status,” *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 27, no. 3, p. 1557, Sep. 2022, doi: 10.11591/ijeecs.v27.i3.pp1557-1566.

[27] A. Ul Haq *et al.*, “Recognition of the parkinson’s disease using a hybrid feature selection approach,” *Journal of Intelligent & Fuzzy Systems*, vol. 39, no. 1, pp. 1319–1339, Jul. 2020, doi: 10.3233/JIFS-200075.

[28] S. Kumar Bhoi *et al.*, “Prediction of diabetes in females of Pima Indian Heritage: a complete supervised learning approach,” *Turkish Journal of Computer and Mathematics Education*, vol. 12, no. 10, pp. 3074–3084, 2021.

[29] J. Ramesh, R. Aburukba, and A. Sagahyroon, “A remote healthcare monitoring framework for diabetes prediction using machine learning,” *Healthcare Technology Letters*, vol. 8, no. 3, pp. 45–57, Jun. 2021, doi: 10.1049/htl2.12010.

[30] A. Perdana, A. Hermawan, and D. Avianto, “Analyze important features of PIMA Indian database for diabetes prediction using KNN,” *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 12, no. 1, pp. 70–75, Mar. 2023, doi: 10.32736/sisfokom.v12i1.1598.

BIOGRAPHIES OF AUTHORS

Dalia Ebrahim Hamid is teaching assistant at college of Electronics and Communication Engineering, Delta University for Sciences and Technology, Egypt. Received a B. Sc degree in electrical engineering from Delta University for Science and Technology, in 2019. She can be contacted at email: Dalia.Hamid@deltauniv.edu.eg.

Hanan M. Amer is associate professor at Mansoura University. She holds M.sc. and B.sc. in Electronics and Communication Engineering from Mansoura University. She has 9 published papers. She is the second supervisor in more than 35 Ph.D. and master students. She can be contacted at email: eng_hanan_2007@mans.edu.eg.

Hossam El-Din Moustafa is currently an Associate Professor with the Department of Electronics and Communications Engineering and the Founder and an Executive Manager of the Biomedical Engineering Program (BME), Faculty of Engineering, Mansoura University. His main research interests include biomedical image and signal processing and deep learning applications. He can be contacted at email: hossam_moustafa@ieee.org.

Hanaa Salem Marie received a bachelor’s degree in electronics engineering from the Faculty of Engineering, Mansoura University, a master’s degree in automatic control system engineering from the Faculty of Engineering, Mansoura University, and a Ph.D. degree in artificial intelligence and image processing from the Computer Science and Engineering Department, Electronic Engineering Faculty, Minufia University. She is interested in data science, big-data analysis, image processing, and machine learning. She can be contacted at email: hana.salem@deltauniv.edu.eg.