AI Based Secure Analytics of Clinical Data in Cloud Environment: Towards Smart Cities and Healthcare

Aghila Rajagopal [1], Sultan Ahmad [2,3,*], Sudan Jha [4], Hikmat A. M. Abdeljaber [5], and Jabeen Nazeer [2]

[1] Department of Computer Science and Business Systems, Sethu Institute of Technology, Pulloor, Kariapatti, India; Email: aghila25481@gmail.com (A.R.)

[2] Department of Computer Science, College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, Alkharj, Saudi Arabia; Email: s.alisher@psau.edu.sa (S.A.), j.hussain@psau.edu.sa (J.N.)

[3] Department of Computer Science and Engineering, University Center for Research and Development (UCRD), Chandigarh University, Punjab, India;

[4] Department of Computer Science and Engineering, School of Engineering, Kathmandu University, Kathmandu, Nepal; Email: sudan.jha@ku.edu.np (S.J.)

[5] Department of Computer Science, Faculty of Information Technology, Applied Science Private University, Amman, Jordan; Email: h_abdeljaber@asu.edu.jo (H.A.M.A.)

*Correspondence: s.alisher@psau.edu.sa (S.A.)

***Abstract*—Recently, health care data comprises of a huge number of information which is regarded as a challenging one for manual maintenance. Owing to the development of big data in the healthcare and biomedical communities, the study of accurate medical data aids the recognition of early-stage disease prediction. As there were several techniques employed for the classification of disease, there were some limitations like low prediction and accuracy rate. To overcome this, deep learning-based classifier is presented. An Artificial intelligence scheme for disease prediction and privacy preservation using Identity based dynamic distributed Honey pot algorithm is proposed in this work for cloud security. Initially, the input medical dataset is preprocessed using normalization technique in which the missing values are replaced and the unwanted data are removed. Whale Optimization based passive clustering is employed for clustering huge data. Multi-scale grasshopper optimization is employed for the process of optimization to get best fitness value. Then the feature extraction and using Robust Shearlet based Feature Extraction algorithm. The classifier is responsible for predicting the disease and for this a Modified Long Short-Term Memory-Convolutional Neural Network (MLSTM-CNN) based classifier is used which provides high accuracy of prediction. Then the data are stored in cloud server or maintenance and monitoring purpose. It is essential to preserve the personal heath record from cloud attack. So as to satisfy this privacy reservation scheme cryptographic techniques are employed in this work. The PHR maintenance is done initially using Identity based dynamic distributed Honey pot algorithm for encryption. Finally, the performance analysis is carried out and the comparative analysis of proposed and existing techniques is done to prove the effectiveness of proposed scheme.**

***Keywords*—artificial intelligence scheme, multi-scale grasshopper optimization, whale optimization based passive clustering, modified Long Short-Term Memory Convolutional Neural Network (LSTM CNN) based classifier, identity based dynamic distributed Honey pot algorithm, cloud server**

I. Introduction

Cloud computing has emerged as a model for an IT company to improve its capabilities on the go without having to invest in new hardware, software, or personnel training [1]. Cloud computing allows you to access a distributed archive rather than a centralized disc unit or proprietary hard drive. It connects and executes the programs as long as the data network has an Internet connection. Individuals and companies alike prefer cloud hosting for a variety of reasons, including cost savings, increased flexibility, speed and reliability, performance, and security. For individuals concerned about data security in the cloud, cloud encryption is essential. Cloud security poses a threat to cloud security. Maintaining data security in the cloud extends beyond the server. Cloud users must safeguard cloud access that can be acquired without caution by saving or login information on mobile devices. Another issue with cloud security is the capacity to protect data collected on a cloud-based server in another country under different laws and privacy standards. They are vulnerable to a variety of risks, including data loss, spoofing, service disruption, wasted electricity, unknown gateways, and so on, due to insufficient preventive measures and a lack of specialized systems for anomaly detection. This can have disastrous consequences, including hardware damage, system access disruption, system blackouts, and even bodily injury to individuals. As a result, the severity of the effects of cloud attacks varies greatly [2, 3]. There are numerous notable technologies

Manuscript received May 11, 2023; revised June 12, 2023; accepted July 19, 2023; published October 26, 2023.

available, one of the most advanced is cloud computing. It provides always-on, appropriate, condemned network access to a community group of convictable computing assets that can be swiftly defend and unconfined with minimal effort and interaction with the service provider [4]. Cloud computing is an integration of numerous computing technical concepts that provides a plethora of business prospects for experts and can meet consumers' needs by storing essential data online [5]. This technology also has a lot of remuneration, which will propel it to the next level [6]. However, due to multiple issues such as loss of self-control and security concerns, cloud computing technology suffers greatly. In cloud computing, the security issue is a big eradicating issue. Because cloud computing is such a novel structure, several users have reservations about it, particularly in terms of security upkeep [7, 8]. The main contribution of this work is to present an Artificial intelligence scheme for disease prediction and privacy preservation using Identity based dynamic distributed Honey pot algorithm for cloud security. The main intention of this proposed work are as follows:

- To present an Artificial intelligence scheme for disease prediction and privacy preservation using Identity based dynamic distributed Honey pot algorithm for cloud security.
- To preprocess the input data using normalization technique to replace missing values and to remove unwanted data.
- To employ Whale Optimization based passive clustering for clustering huge data.
- To employ multi-scale grasshopper optimization for the process of optimization to get best fitness value. To extract features using Robust Shearlet algorithm.
- To predict and classify the disease using Modified Long Short-Term Memory-Convolutional Neural Network (MLSTM-CNN) based classifier which provides high accuracy of prediction.
- To store the data in cloud server for maintenance and monitoring purpose.
- To preserve the personal heath record from cloud attack and to satisfy privacy reservation using cryptographic techniques.

The following is how the paper is structured. The necessity for data integrity techniques and cloud storage are discussed in Section II as heading related works. Section III takes a thematic look at privacy-preserving cloud storage solutions. Section IV demonstrated cloud security techniques. The conclusion and Future research directions in cloud security techniques were provided in Section V.

II. Related Works

Cloud computing is envisioned as the foundation for the next generation of IT initiatives. It moves the software programmes and database to big data centers in the cloud, where data and service management may not be completely reliable [9, 10]. This unique methodology uncovers various new security challenges that were previously unknown. The data storage integrity issues and solutions in Cloud Computing environment has been proposed in two different recent works [11, 12]. Allowing a Third-Party Auditor (TPA) task to certify the active data integrity maintained in the cloud is specifically explored in support of the cloud client [13]. The introduction of TPA eliminates the client's participation in the evaluation of his data stored in the cloud, which can be significant in achieving scale economies for the cloud [14]. The data dynamics are carried through the most general data operation forms, such as deletion, insertion, and block modification, and are also a crucial stage to practicality in cloud computing services. Previously published research on establishing distant data integrity usually lacked public auditability and dynamic data support operations [15]. The potential security issues and difficulties of direct extensions were identified first with the entire informs of dynamic data from previous works, and then demonstrated how to build an elegant authentication system for the seamless integration of these two important features in the protocol design. In particular, we can enhance the current storage model verification on the classic Merkle Hash Tree (MHT) building deployment for block tag verification to achieve effective data dynamics. To aid in the efficient management of numerous auditing activities, the technique of bilinear aggregate signature was investigated to cover our main result in a multi-user situation, where TPA can perform many auditing tasks at the same time [16]. The proposed systems are exceptionally effective and provably protected, according to extensive performance analysis and security [17].

Users can store their data remotely and take advantage of high-quality applications with cloud storage [18]. The work [19] presented a paper that focuses on Cloud integrity and privacy algorithms that rely on hardware tamper-proof and energy-efficient cryptographic data structures. This in turn examines two critical design patterns that go across every successful BSN exploitation and are denoted by:

- Finding the right balance between the degree, strength, span, and complexity of the cryptographic functions in use and the energy resources consumed.
- Achieving a possible trade between the human subject's confidentially and the subject's safety while wearing the BSN. This is followed by an analysis of practical cryptographic support in the leading BSN development frameworks, including SPINE and TinyOS, which leads to a collection of recommendations and generic guideline patterns for implementing and designing cryptographic procedures in the BSN framework.

Yadav *et al*. [20] proposed an effective public integrity auditing approach with user revocation by a comfortable workforce based on commitment vector and revocation verifier-local group signature. A concrete scheme is established by a novel structure known as Decrypt key, which provides a reliable and effective declaration for convergent key management on common user with cloud storage sides.

In another work, Fan *et al*. [21] suggested an identity-based protected combined signatures (SIBAS) approach for data integrity verification that uses the Trusted Execution Environment (TEE) as the auditor to examine outsourced data on the local side. SIBAS can not only verify the integrity of outsourced data, but also maintain secure keys in TEE via Shamir's threshold scheme.

Zhang *et al.* [22] proposed a unique public verification strategy for cloud storage based on indistinguishability complication, which requires an insignificant calculation on the auditor and the envoying of the most computation to the cloud. A strategy to facilitate batch verification and data dynamic operations was also presented, in which the auditor may efficiently execute numerous authentication duties from different users and the data of cloud-stored ones can be dynamically updated. When compared to previous existing works, this system greatly minimizes the auditor's computing overhead. Furthermore, the auditor's overhead on batch verification is independent of the amount of verification tasks in this method.

Imran *et al*. [23] proposed a technique to handle the data integrity problem in Cloud computing by allowing users to validate the Cloud data integrity saved. Users can also track data integrity violations if they occur. In Cloud computing, a new notion called "Provenance of Data" was used to help with this durability. This solution can reduce the need for third-party services while also supporting hardware and data item replication on the user side to ensure data integrity. Hasan *et al*. [24] presented a novel resource oriented DMA framework for internet of medical things devices in 5G network. A lightweight cryptographic technique for the purpose of guessing attack protection in the complex IoT application was presented in Hasan *et al*. [25]. In another work [26], the issues of clinical identity verification for healthcare applications over mobile terminal platform in cloud environment has been discussed.

In order to overcome limitations in the existing approaches like data integrity problem, reduced classification accuracy and less security an effective technique is presented in this work.

The following are the primary contributions of the suggested technique:

- To provide a passive clustering technique for medical data based on whale optimization.
- Using the Multi-scale grasshopper optimization technique, optimize the clustered data.
- Robust Shearlet-based feature extraction technique to extract features.
- Using the Modified LSTM-CNN classifier, categorise the input data as normal or abnormal and forecast diabetic illness.
- Encrypt the predicted data with an Identity-based dynamic distributed honey pot technique and store it in the cloud, where it may be decrypted as needed by the user.

III. Proposed Work

This section delivers a detailed explanation of the proposed system. The overall flow of the proposed system is shown in the Fig. 1.

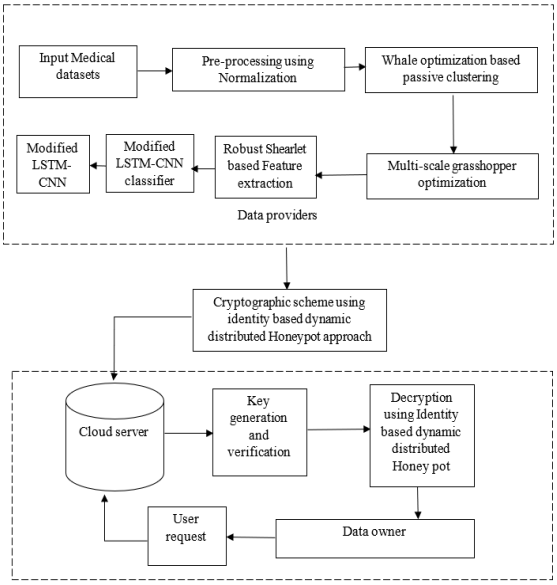

Figure 1. Flow of the proposed system.

A. Preprocessing Using Normalization

The input KDD cup IDS dataset is normalized for preprocessing, which is the first step in the suggested technique. For a more pleasant design and more productive use, the pre-processing dataset is proposed to the raw data transformation for future processing steps. (1). Data alteration (encoding); and (2). Normalization are included in the preparation stage. When some feature values are never evident for the Neural Network throughout the data alteration (encoding) step (NN). They should be swapped for characteristics in the proposed method, which is called Encoding, with detectable amounts.

Both the training and test data sets are first homogenized. The weight coefficients selection approach relies heavily on data normalization. The detection rate is not manipulated in the same way that the Normalization technique is. The normalized new value is calculated using the equation below. The new value's mapping range is 0 to 1. Because all of the data used in training has a comparable field, such as between 0 and 1, normalization can improve the training instance.

$$Y_{\text{norm}} = \frac{Y - Y_{min}}{Y_{max} - Y_{min}} \quad (1)$$

where Y_{norm} is the normalizing result and Y is the original value before normalization. Y_{max} and Y_{min} represent the maximum and minimum values of each feature, respectively.

The primary goal of data normalization is to minimize or eliminate redundant data. The dataset is normalized in terms of missing data achievement and the input IDS datasets are also normalized by considering this preprocessing stage.

B. Whale Optimization Based Passive Clustering

This segment uses a clustering approach to collect data from multiple sectors. Clustering and WO technologies are also combined to improve overall efficiency. For a

conventional *N* input dataset, the data is divided into *L* clusters. The total number of data in each cluster is [N / L]. Second, the region partition line is computed using the passive clustering algorithm WO, dividing the data into two regions.

$$N = \alpha\, Ko + \beta\, Kn,\ \alpha, \beta \in [0,1],\ \alpha > \beta. \quad (2)$$

where α, β is the horizontal and vertical coordinates point line segmentation, *Ko* and *Kn* is the angle between the line and axis. After that the fitness value *k* and *n* particles are calculated as below:

$$F(k,n)=[f(k)L(k/n=k(k)/\int_{-\infty}^{\infty} f(k)(u)L(k/n = y(k)(u)\mathrm{du}] \quad (3)$$

where *F*(k,n) is the fitness value function, *f*(*r*) *L*(*k*/*n*=*n*(*k*) which is a likelihood function, $f(k)(u)\mathrm{L}(r/n) = n(k)(u)$ which is a normalizing constant.

Because whales have cells that are equivalent to those in the human brain, the WO algorithm offers a novel optimization tool. One of the well-known creatures is the whale. The technique, like other optimization procedures, begins with random solutions of individual parameters before determining the best goal function. Every iteration in WO, the search agent updated the location, and the target function was defined based on this modification. The loop is repeated until most iterations respond, at which point you compose the optimal response.

Surrounding beast whales, bumper whales, and humpback whales are the three steps of the WO optimization process. The bumper whales first recognize the beast and then circle it. First and foremost, the WO approximation means that the optimal result is greater than the maximum. Other whales try to better alter their posture if the reaction is the strongest. The Whale Optimization Method concept is a metaheuristic algorithm for solving optimization problems and performing a mathematical derivation test for the global optimum. The Whale Optimization Algorithm (WOA) was created using whale hunting techniques. The following method is a bubble net feeding technique. Humpback whales catch small fish below the surface by letting a net of bubbles expand in a circular pattern around the prey. The following is a mathematical representation of the whales' prey methodology:

$$W_{surrounding}=\left|\overrightarrow{N}\overrightarrow{y \times (T)} - \overrightarrow{\vec{y}}\,(T)\right| \quad (4)$$

$$\mathrm{y(T + 1)} = \vec{y} \times \mathrm{(T)} - \vec{A}\mathrm{P} \quad (5)$$

where *T* is the iteration, $\vec{A}$is the vectors coefficient, $\vec{y}$* signifies vector of best solution, $\vec{y}$ indicates the position vector, *N* denotes the absolute value, and '·' is the indication of a multiplication element-by-element.

The best solution's position vector is updated once an improved solution is recognized.

$$\vec{\alpha} = 2\vec{\alpha}\ \vec{s} - \vec{\alpha} \quad (6)$$

$$\overrightarrow{N}=2\vec{s} \quad (7)$$

where is lowered from 2 to 0 linearly over the number of iterations (in stages of exploitation and research) in Eqs. (6) and (7), and is an arbitrary vector in the range [0, 1]. This limitation is changed in order to preserve the balance between the exploitation and analysis stages. The whale and prey locations are balanced here; whales follow each other at random based on their roles.

With random numbers more than one or less than one, the search agent will push further away from the reference whale. At the inquiry level, the search agent status was modified by a randomly selected search agent rather than the strongest search agency available. The quest phase is based on the vector fluctuation and is done manually. Humpback whales are blindly hunting for the optimal place based on one another. The mathematical expression for the best global position is as follows:

$$P = |\vec{\alpha}|\vec{Y} - \vec{Y}_{rand} \quad (8)$$

From that, the optimal solutions were obtained. Where $\vec{Y}_{rand}$ is a random whale position vector selected from the current population. The Minimum function value is usually written as an Eq. (9),

$$\frac{\alpha(s)}{e(s)} = N_p{+}N_i/s{=}N_p\left(1 + {}^{1}/_{k_i}\, s\right) \quad (9)$$

Finally, the minimum value is reached. With the necessary particles, it can be utilized as an extreme global value. Similarly, a single particle is assigned the lowest fitness value. The values are then adjusted to the fitness *f's* minimum value. The two subsections are split after the region is segmented until the final *N* clusters are generated. The cluster head can then be chosen. The correlation groups are then formed after spatial adjustment. Spatial analysis provides new views for your decision-making. Each spatial correction term for each cluster n should, on the surface, be changed separately. To put it another way, a *i* is better for the total parameter *j*, which is linked to each of the *n* clusters. In this, the subsequent modified objective function was regarded as,

$$D_{an} = \sum_{l=1}^{l=m-1} Oj(k_l\,, k_{l+1})\frac{1}{m} \cdot (m-1) \quad (10)$$

where k_l is the information for the data carrier, *l* is the number of data carrier used by the entire time domain, and $Oj(k_l, k_l +1)$ is the system link between k_l and $k_l +1$. The data in the system is S, and the contact time between k_l and k_l+1 is OA(k_l, k_l+1). *D* here is k_l to the central field of concern, the distance between the current data carrier and the next data must first be taken into account.

The distance between was computed and the cluster head was elected by selecting the cluster members. Here the correlations have to be performed periodically.

$$C\left(\frac{p}{\partial}, \mu\right) = (\mathrm{X(correlated(ji))} + \mathrm{e})\left[\frac{\vartheta(\partial+\mu)}{\vartheta(\partial)\varphi(\mu)}\right] p\mathrm{\wedge}(\partial + \mu)\mathrm{\wedge}(\mu - 1) \quad (11)$$

where ∂, p, μ represents the correlated features of data, ϑrepresents the spatial variance. As a result of it correlated groups get formed.

Then by characterizing data we will estimate the transmission probability. Let us denote for all $t \geq 0$ and all $i \in K$,

$$\propto_t{}^i = \mu_t(\mathrm{i}) / \sum_{j \in K} \mu_t(\mathrm{j}) \quad (12)$$

$$\beta_{\emptyset t} = \mu_t(\mathrm{i})\text{^}2 / \sum_{j \in K} \mu_t(\mathrm{j})\text{^}2 \quad (13)$$

The probability measures αt and $\beta_{\emptyset t}$ have instinctive elucidations: The first one describes the distributions of a degree randomly (and uniformly) unexplored data picked at t time, while the next one is the size influenced αt distribution and signifies the distribution degree of all data of an unexplored input that were picked randomly or else, the degree of starting data at a half-edge uniformly drawn at random, between all half-edges opening from unfamiliar data. Let us estimate probability. Hence the transmission probability can be calculated by using the formula,

$$\mathrm{S}(\emptyset / z) = (\mathrm{p}\frac{(\frac{q}{\emptyset})}{\mathrm{p(q)}})\mathrm{P}(\emptyset) \quad (14)$$

where $s(\emptyset)$ is the probability distribution function, $P(\frac{q}{\emptyset})$ is the likelihood function, $P(q/z)$ is the evidence function,

After that the allocation of fractional part of time slots to the member of cluster groups by the transmission probability.

$$\text{Energy estimation} = E_{e_{ch}} + 1 \quad (15)$$

Then after allocation of time slot the optimal cluster path can be calculated using the multi-scale grasshopper optimization.

C. *Multi-scale Grasshopper Optimization for the Optimization of Clustered Data*

Multi-scale resources are usually The Grasshopper optimization algorithm (multi-scale resource GOA) can produce better results in a fair amount of time. This hypothesized mechanism's exploitation, convergence, and exploration abilities are all examined.

The maximum number of iterations used in this optimization procedure is 1,500. Maximum iteration, search agent number, lower bound and upper bound dimensions, and predicted distance are the parameters to consider (x-coordinate, y-coordinate, clustered data). The optimization procedure then proceeds by estimating the fitness function, which then determines the efficiency. The fitness function evaluation is carried out in order to determine the ideal values for closeness or distance in order to increase efficiency. The procedure shown below shows how to calculate the optimal fitness function value. As a result, the multi-resource grasshopper optimization process is used to carry out the optimization technique. Along with the random population of grasshoppers, the parameters are being initialized. Closeness probability degree and Gaussian probability degree are used to estimate the multi-objective function. The individual grasshopper positions are then updated. The best position is reached when the population size surpasses the number of iterations, and the current best position is designated as the old best position. If the condition is not met, the process is repeated until the desired result is achieved. This procedure's algorithm is depicted below as Algorithm 1.

Algorithm 1. Multi scale resource Grasshopper Optimization

Input: clustered data d_u

Output: optimized_valueO_{val}

Step1: initialize the parameters,
No of clusters b_s
Step:2 optimization
Search_agenta_s=N_u
Max_iterationi_{opt}=500
Lower_bandb_L=-100
upper_bandb_U=100
dimensions μ =b_s
cmax=1
cmin=0.00004
while 1<i_{opt} + 1
c=cmax−1$(cmax - cmin)/i_{opt}$
for loop
update GH position
To find distance
New GH positions
Goal positions
end
step 5:
Optimization function
O_{val}=b_U × exp (−μ/1)-exp(-b_U)

The distance user is the input in the Multi-scale resource Grasshopper optimization approach. The settings for this are being set up at the moment. The search agent is then examined in the optimization technique with a maximum number of iterations of 500, as well as lower and upper bounds of 100 and dimension. The upper and lower bound limits are fixed, and if the condition is met, the GH position is updated using distance, new GH positions, and global positions. Finally, the best fitness function and the best cluster data are found.

D. *Robust Shearlet Based Feature Extraction*

Following the optimization step, the features were picked using the adaptive Shearlet method (ASA). This is a technique for deleting second-order mathematical texture features. Higher-order features are used to interact with three or more data properties, and this technique has been used in various applications. This is a mathematical task that will usually efficiently delete the unneeded data. The data's correctness can also be displayed plainly. During the analysis cycle, the data is distinguished. Shearlet can indicate the frequency of data attributes in a specific exact differential feature. Another pixel is recognized as the route l and the neighboring value detachment of m, and the single-data is questioned here. m usually gets a single value and can benefit in both directions. The obtained directional value can remove the attributes of the data used for the

classification process. The Shearlet process may be set as follows:

$$K(l,n)=G(l,n,o,\emptyset)/\sum_{l=1}^{H}\sum_{n=1}^{H} G(\ l,n,o,\emptyset) \quad (16)$$

where G represents the frequency vector, l, n, o represents the frequency of a certain component with pixel values of 0 and 1, K represents the data features, (l, n) represents the component of the l, and l represents the normalized constant.

On the database, the ASA uses N-dimensional vectors to represent the position of individual components throughout the feature space. The characteristics can then be selected based on their relationship. The method evaluates the subset by considering the predictive potential and redundancy of each function individually (or similarity). This means that the algorithm will choose the option that improves the performance of this feature by using a (heuristic) function to determine its future steps. Heuristic functions can also be used to solve problems. By using the following equation, the correlation between the features can be defined

$$K\left(\frac{o}{\partial},\mu\right)=\left[\frac{\varphi(\partial+\mu)}{\varphi(\partial)\varphi(\mu)}\right] o\hat{}(\partial+\mu)\hat{}(\mu-1) \quad (17)$$

This method helps to process the data and extracting the features of the affected region from the input data in an effective manner.

E. Modified LSTM CNN Based Classifier for Classification and Prediction of Disease

Unlike a traditional neural network, the secret layer's basic unit is a memory cube. The memory block contains time-saving memory cells that are monitored by a pair of adaptive multiple gating devices. The source of input and output controls the block's activation input and output in two ways. Activating memory cells can be done first. As a result, the proposed correlation feature-based LSTM can tackle the fading error problem while keeping the error constant. The specified features can be offered as an input to the procedure after the memory block has been activated. As the flow of information becomes redundant and the error input weight is replaced by the weight source activation, the memory blocks are free to reset themselves. The proposed classifier's main principle is a forward and backward training sequence to two unique recurring systems that are both related to the same performance stage. This means that combining all points before and after each point in a series yields the most sequential knowledge.

Step 1: Create an N-point random training set.

Step 2: Based on the N points, a decision tree can be constructed.

Step 3: Select the exact replicate from phases 1 and 2 for use in constructing a decision tree.

Step 4: Create a unified probability based trusted value.

A collection of input datasets is generated by the Modified LSTM CNN that can be shown in Algorithm 2. The final test entity category is determined by combining the votes of classifier subsections into a single word. The Unified probability with LSTM neural network for predicting the dangerous disease in the input data must be determined in order to calculate the trustworthy classification value. Here you can calculate the Unified probability using LSTM neural network parameters.

Algorithm 2. (Modified LSTM CNN classification)

Input: Specialized features S_f
Output: Classified dataC_{datas}
Initialize the multi-Network layers
Initialize train features
Initialize label
Train label =80%
Tet label =20%
Label=unique(label)
For ii=1:lengh(Lab)
  Class=find(label== Lab (ii))
Traincut=length(class)-traincut
Traindata=[traindata; trainfeatures; class(1: Traincut)end-5:end]
  Predict label=classify(net,traindata)
End
End
For ii=1:size(traindata,1)
Traindata=[traindata; trainfeatures;class(1: Traincut)end-5:end]
End
For ii=1:size(trainfeatures,1)
Traindata=[trainfeatures; trainfeatures;class(1: Traincut)end-5:end]
End

$$C_f(i)=\begin{cases} c_{f\,min}+\left(c_{f_{max}}-c_{f_{min}}\right)\times n, n<1 \\ c_{f_{max}}/N_{iter} \end{cases}, i=1, \ldots, n \quad (18)$$

where f = fitness(i) − f_{min}, which depends on the current quality of i_{th} solution

After that the parameter gets initialized and the abnormality gets analyzed.

$$\sigma_v=(\gamma(1+\beta_c))\times\sin(\pi\times\beta_c/2)/(\gamma(1+\beta_c\ 2)\times{\beta_c}^{\beta_c-1/2}) \times(\beta_c-1/2)^{1/(\beta_p-1/2)} \quad (19)$$

where σ_v represents the random size of the nest.

$$\text{Correlation }(i,j)=\sqrt{(C_{n_{fea}}(i,1)-V_{n_{fea}}(j-1)} +C_{n_{fea}}(i,1)-C_{n_{fea}}(j,1)^2 \quad (20)$$

As we conditioned the prototypes on normal and abnormal values depend upon its abnormal features. Here the probability can be calculated by using the Eq. (21).

$$C_{n\_prob}=[C_{n\_prob}\text{unif}] \quad (21)$$

Hence the classifier can evaluate the probability of abnormality and can discriminate the abnormal state. Then the trust value can be calculated for the classification of the input data.

$$C_{tv}=(C_{n\_prob}dist) \quad (22)$$

where C_{tv} represents the trusted value of the input data. Depends upon the trusted value the abnormality can be identified.

By concerning these processes, to attain the score value, which is effectively choose either the particular data is normal or not. If the score value surpasses the threshold value, it denotes that the precise interference, i.e., normal data. If the threshold value is lesser than or equal, the particular data considered as the abnormal condition.

F. Encryption of Predicted Result Using Identity Based Dynamic Distributed Honeypot Algorithm & Cloud Storage

Honeypot is an energetic protection method in which reserves are placed in a system with the goal of monitoring and limiting initial attacks. This research presents a Honeypot-based Intrusion Detection System (IDS) prediction in order to make better use of facts about the attacker. Such extraordinary issues require a very robust Honeypot-based prediction technique to identify intrusion detection-based attacks. Honeypots are used to defend the system. A honeypot is a critical security entity that gives up its benefit to unauthorised admittances in order to compute potential flaws in operational structures and eliminate hazardous/dangerous situations. These are catches for the unwary user. Honeypot is a mechanism that is set up to attract attackers. Honeypots are used in a variety of fields, including intrusion detection systems. It holds the False Negative Rate (FNR) and False Positive Rate (FPR) when used with Intrusion Detection System (IDS) and firewall (FPR). It also adds a layer of security-related information. It also works effectively with encryption and communication during IPV6, unlike other security measures. Algorithm 3 shows the process steps to be follow in this method.

Algorithm 3. Identity based Dynamic Distributed Honeypot (IDDH) algorithm for encryption

```
Input: Intrusion History Database
Output: Revised inspection vector Curr O and P(O)
Begin;
while Intrusion is inferred, do
  Choose similar intrusions from IHD;
if the Possibility that intrusion persists is greater than intrusion discontinuing then
    Honeypot permits the intrusion;
    Reacts with a reply;
else
    Honeypot obstructs the intrusion;
    Falls the message of the intrusion;
```

The goal of the proposed study is to forecast incursions in the system by first establishing feature extraction and feature selection. Then, based on the features, create a data clustering. The intrusion formation is achieved via a series of clusters in which the incursions are present. The incursions may or may not be present in each cluster group. The clustering method used in Honeypot-based intrusion prediction. According to the interaction level of the Honeypots, the Honeypot-based prediction approach aids in identifying the intrusions present in the cluster formation. The steps involved in the proposed task are listed below. Honeypots are typically classed as High-interaction, Less-interaction, or Medium-interaction based on the level of engagement between the intruder and the system.

They either follow a complete operating system or use a genuine operating system installation with a supplementary monitoring system at the High-interaction level. They not only manage requests, but they also allow harmful schemes to totally intertwine while still allowing the advised technique to be used. Several high-interaction honeypots additionally allow for limited peripheral connections, allowing the service to remain operational during DoS attacks while preventing it from receiving data, resulting in huge traffic and losses.

In the Less-interaction level, an intruder can monitor in but not execute any actions since the system is examined rather than the full system. This is a very safe resolution that poses little risk to the environment in which it is implemented. The reduced interaction Honeypot exclusively uses the system's transport layer on a single physical host and never uses data gathering on the application layer. Attackers never contact with the genuine Operating System (OS), instead gaining access to imitation services such as a fake web or mail server. However, the data acquired by the Honeypot with minimal input can reveal vital details about the intrusion, such as if an attacker attempted to exploit well-known vulnerabilities in certain services. Honeypots with less interaction evolve.

1. They forecast and detect attacks.
2. The ability to detect genuine login attempts, as opposed to passive IDS systems.

Medium-interaction Honeypots, which are a hybrid of low- and high-interaction honeypots, are capable of tracking entire services or specific vulnerabilities. Its major goal, like that of less-interaction Honeypots, is prediction, and it is set up as a purpose on the host OS, with only the imitation services available to the public.

Honeypot is used in the proposed work to determine an unauthorized exploit of an information system by studying the attacker's behavior in an isolated and monitored environment. They can be cluster-based systems, but they're more common, and network-based contact is never performed by a network association. The primary benefit of a honeypot is that it simplifies the Intrusion Detection problem in terms of intruded or non-intruded networks by requiring no genuine use. As a result, any movement on a Honeypot can be classified as odd right away. Honeypots are well-suited to monitoring harmful network movement due to their properties. Honeypots are designed to imitate systems that an attacker might want to break into, but they prevent the invader from gaining access to the entire facility. Honeypot either allows or prevents the assailant from carrying out their actions. Honeypot removes the Associate's Position, AP(Qm), from each Qm in Curr O to pick the response to the attacker. The term of AP(Qm) as given below:

$$\mathrm{AP}(Qm) = \{Oi/Oi \in O/Or \,\&\{OiOr\} \in \boldsymbol{\phi}(Ti)\} \quad (23)$$

The Associate's Position of Or, AP(*Qm*), is a collection of observations that explain any hint of *Ti* when *Or* is present. Honeypot responds to every *Oj* in AS(*Ok*). Honeypot either responds with the next message in the sequence specified by the protocol's standards, or it refuses to allow the process to expand further. The Honeypot selects the achievement by submitting it to the IHD (Intrusion History Database), where the IHD (Intrusion History Database) is a database of network intrusions.

The intrusion was predicted based on the interaction level. If a honeypot is correctly predicted, the intruder will have no scheme to defraud and observe. The intrusion is projected based on the interaction level (less, high, medium), and the intruded and non-intruded networks are calculated. The forecast accuracy (how accurately the intrusion was recognized) is examined when evaluating the incursion. The data is saved in the cloud environment if there is no intrusion in the network. Honeypots provide an effective solution to improve network security in terms of non-intruded networks and network consistency in the suggested system. Furthermore, the data is stored in a secure manner with no breakage; security is the most important component in data storage.

IV. Performance Analysis

This section is the detailed explanation of the performance analysis of proposed intelligent based fuzzy system.

A. Dataset Description

Pima Indian Diabetes Dataset provides details from a community near Phoenix, Arizona, the USA on 768 patients (268 tested positive cases and 500 tested negative cases). The Tested positive and Tested negative show whether or not the individual has a diabetic substance. Every instance consists of 8 integer attributes. These data contain personal health information and medical test results. The data set includes the following detailed attributes,

- Number of times pregnant (preg)
- Plasma glucose concentration at 2 h in an oral glucose tolerance test (plas)
- Diastolic blood pressure (pres)
- Triceps skinfold thickness (skin)
- 2-h serum insulin (insu)
- Body mass index (bmi)
- Diabetes pedigree function (pedi)
- Age (age)
- Class variable (class)

In the column of pregnancies, the patient's number of times pregnancies are given. In the column of Glucose, Plasma glucose concentration a 2-h in an oral glucose tolerance test is included. In the blood pressure column, Diastolic blood pressure (mm Hg) is given. In the skin thickness, Triceps skin fold thickness (mm) is given. In the insulin column, 2-h serum insulin (mu U/ml) is included. In the BMI column, Body mass index (weight in kg/ (height in m) ^2) is given. In the outcome column, Class variable (0 or 1) 268 of 768 is 1, others are 0.

B. Performance Metrics

1) Accuracy

It is a measure of arithmetical predisposition. The accuracy is the proportion of true results (both true positive and true negative) in the total data.

$$\text{Accuracy (A)} = \frac{\text{TP+TN}}{\text{TP + TN + FP + FN}} \tag{24}$$

2) Precision

The precision can be calculated by using the Eq. (25),

$$\text{Precision} = \frac{\text{TP}}{\text{TP + FP}} \tag{25}$$

3) Sensitivity

Sensitivity is also a true optimistic rate, the detection possibility, recall in some fields it is regarded as the actual positives measure proportion that were recognized correctly.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP + FN}} \tag{26}$$

4) Specificity

Sensitivity is also a true optimistic rate, the detection possibility, recall in some fields it is regarded as the actual positives measure proportion that were recognized correctly.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN + FP}} \tag{27}$$

C. Performance and Comparative Analysis of Proposed and Existing Techniques

The presented technique is related with the existing state-of-the-art algorithms like Naive Bayes, KNN, Decision Tree, Logistic Regression and Modified SVM [27, 28]. Fig. 2 illustrates the measure of accuracy comparison with the existing methods. The algorithm KNN provides 97.065% value, Naive Bayes offers 96.003% as an accuracy value. Similarly, the decision tree offers 96.628% and Logistic Regression offers 96.92%. The MSVM technique has accuracy value of 97.13%. Whereas, the presented technique offers an accuracy rate of 98.99%. On comparing existing techniques, the presented technique offers better outcome.

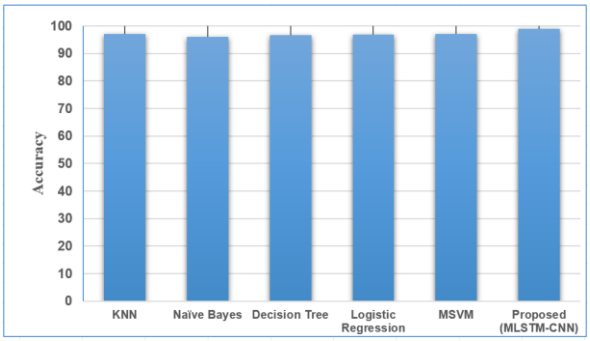

Figure 2. Accuracy (%) comparison with existing methods.

Fig. 3 illustrates the measure of sensitivity comparison with the existing methods. The proposed method is compared with the existing methods like Naive Bayes, KNN, Decision Tree, Logistic Regression and Modified SVM. The algorithm KNN provides 96.838% value, Naive Bayes offers 93.389% as a sensitivity value. Similarly, the decision tree offers 96.171% and Logistic Regression offers 97.298%. The MSVM technique has sensitivity value of 97.458%. Whereas, the presented technique offers a sensitivity rate of 98.04%. On comparing existing techniques, the presented technique offers better outcome.

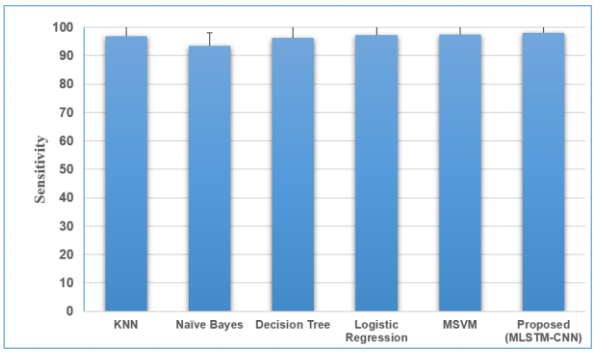

Figure 3. Comparison of sensitivity (%) with existing methods.

Fig. 4 represents the comparative analysis of specificity rate with existing methods. the comparison of specificity measure with the existing algorithms. The algorithm KNN provides 97.484% value, Naive Bayes offers 97.41% as a specificity value. Similarly, the decision tree offers 97.5% and Logistic Regression offers 96.232%. The MSVM technique has specificity value of 97.111%. Whereas, the presented technique offers a specificity rate of 98.20%. The proposed specificity rate is better than others.

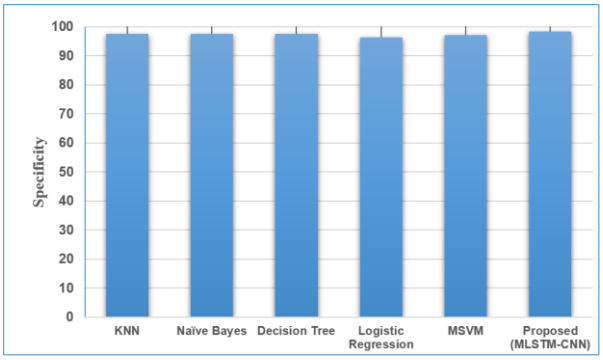

Figure 4. Comparison of Specificity (%) with traditional techniques.

Fig. 5 illustrates the measure of precision comparison with the existing methods. The proposed method is compared with the existing methods like Naive Bayes, KNN, Decision Tree, Logistic Regression and Modified SVM. The algorithm KNN provides 98.646% value, Naive Bayes offers 95.1% as a precision value. Similarly, the decision tree offers 98.699% and Logistic Regression offers 98.017%. The MSVM technique has precision value of 99.333%. Whereas, the presented technique offers a precision rate of 99.861%. On comparing existing techniques, the presented technique offers better outcome.

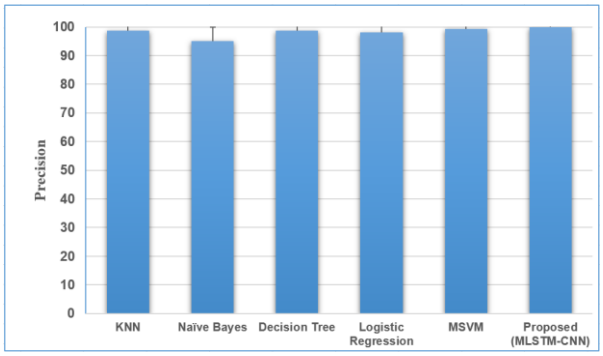

Figure 5: Precision (%) comparison with existing methods.

Fig. 6 depicts the encryption time comparison with that of the existing techniques like RSA, DES, AES, Enhanced honey bee algorithm. The values of encryption time are related with that of RSA, DES, AES, Enhanced honey bee and the proposed algorithm, having the size of 25 KB, 50 KB, 1 MB, 2 MB, and 3 MB correspondingly. This statistical outcome represents that the presented technique consumes less time for encrypting the data in the cloud.

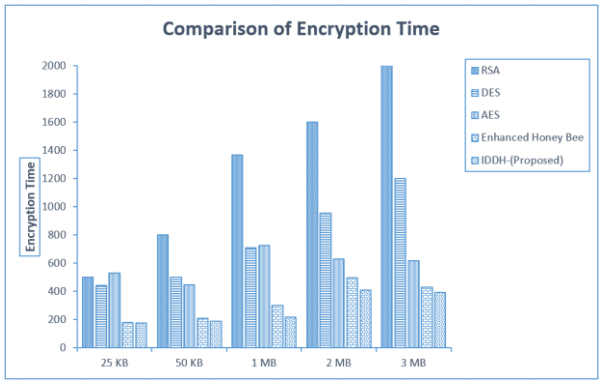

Figure 6. Encryption time (s) comparison with existing methods.

The decryption time comparison with that of the existing techniques like RSA, DES, AES, Enhanced honey bee algorithm is shown in Fig. 7. The values of decryption time are related with that of RSA, DES, AES, Enhanced honey bee and the proposed algorithm, having the size of 25 KB, 50 KB, 1 MB, 2 MB, and 3 MB correspondingly. This statistical outcome represents that the presented technique consumes less time for encrypting the data in the cloud.

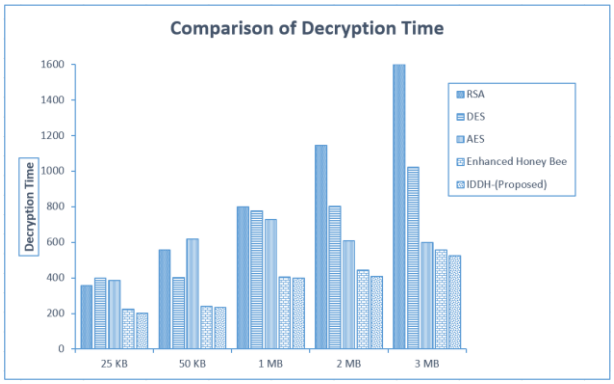

Figure 7. Decryption time (s) comparison with existing methods.

A Hamming loss is estimated and the outcomes attained are projected and are compared with traditional models to compare the efficiency of proposed model with traditional ones [29].

From Fig. 8, it was revealed that the proposed model shows lower hamming loss than other existing models. Therefore, the proposed model offers enhanced outcome.

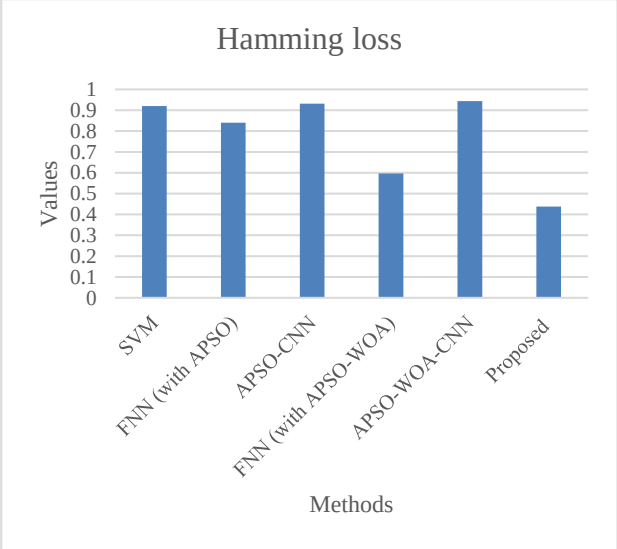

Figure 8. Hamming loss estimation.

V. Conclusion

An Artificial intelligence based MLSTM-CNN classifier system is used to classify diabetes data. This inquiry's fundamental aim is to enable successful and early detection of diabetes and to store the predicted information in the cloud securely. The model proposed consisted of both the prediction and the storage of cloud data. The public PIMA Diabetes data collection has been analyzed. It has achieved the highest predictive accuracy of 98.99%. Here, for assessing the cloud data security, the proposed Identity based dynamic distributed honeypot algorithm was compared with the existing security framework. The performance clearly shows that the proposed technique effectively provides high accuracy and better data security. This patient knowledge can be stored for further data analysis and layout improvement techniques in a database. It allows us not only to consider their wellbeing but also to establish healthier living conditions. The proposed model limitation is the processing time, this is because of huge volume data taken for the estimation of training data performance. In future, this work will be extended by implementing this in real-time data to estimate the efficiency of presented system.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization, A.R., and S. A.; methodology, S. J., H. A. A.; software, H. A. A. and J. N.; validation, S. A, A. R., and S.J; writing–original draft preparation, A.R., S.A and S. J.; writing–review and editing, S.J, J.N.; visualization, H.A. A. and J.N. ; supervision, A.R. and S.J.; project administration, S. J., S.A.; funding, S.A., J.N., H.A.A.; All authors have read and agreed to the published this version of the manuscript.

Funding

This study is supported via funding from Prince Sattam Bin Abdulaziz University project number (PSAU/2023/R/1444).

Acknowledgment

We thank the Prince Sattam Bin Abdulaziz University, Alkharj, Saudi Arabia for help and support.

References

[1] L. Zhou, A. Fu, S. Yu, M. Su, and B. Kuang, "Data integrity verification of the outsourced big data in the cloud environment: A survey," *Journal of Network and Computer Applications*, vol. 122, pp. 1–15, 2018.

[2] H. Karajeh, M. Maqableh, and R. Masa'deh, "Privacy and security issues of cloud computing environment," in *Proc. the 23rd IBIMA Conference Vision*, 2020, pp. 1–15.

[3] S. Ahmad and M. M. Afzal, "A study and survey of security and privacy issues in cloud computing," *International Journal of Engineering Research&Technology (IJERT)*, vol. 6, no. 1, pp. 429–432, 2017. https://doi.org/10.17577/ijertv6is010311

[4] R. Hariharan, G. Komarasamy, and S. D. M. Raja, "An extensive review on data integrity schemes and security issues in cloud paradigm," *International Journal of Advanced Research*, vol. 8, no. 6, pp. 1093–1100, 2020. http://dx.doi.org/10.21474/IJAR01/11192

[5] M. Y. Uddin and S. Ahmad, "A review on edge to cloud: Paradigm shift from large data centers to small centers of data everywhere," in *Proc. 2020 International Conference on Inventive Computation Technologies (ICICT)*, 2020, pp. 318–322. doi: 10.1109/ICICT48043.2020.9112457

[6] S. Jha, S. Ahmad, M. Alharbi, B. Alouffi, and S. Sebastian, "Secured and provisioned access authentication using subscribed user identity in federated clouds," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 12, no. 11, pp. 178–187, 2021.

[7] D. G. Rosado, R. Gómez, D. Mellado, and E. Fernández-Medina, "Security analysis in the migration to cloud environments," *Future Internet*, vol. 4, pp. 469–487, 2012.

[8] S. Jha, D. Prashar, H. V. Long, and D.Taniar. "Recurrent neural network for detecting malware," *Computer Security*, vol. 99, 102037, 2020.

[9] H. Tian, F. Nan, C. C. Chang, Y. Huang, J. Lu *et al.*, "Privacy-preserving public auditing for secure data storage in fog-to-cloud computing," *Journal of Network and Computer Applications*, vol. 127, pp. 59–69, 2019.

[10] S. Jha, D. Prashar, and A. A Elngar, "A novel approach using Modified Filtering Algorithm (MFA) for effective completion of cloud tasks," *Journal of Intelligent & Fuzzy Systems*, vol. 39, no. 6, pp. 8409–8417, 2020.

[11] C. Wang, S. S. M. Chow, Q. Wang, K. Ren, and W. Lou, "Privacy-preserving public auditing for secure cloud storage," *IEEE Transactions on Computers*, vol. 62, no. 2, pp. 362–375, Feb. 2013.

[12] S. Jha, L. Nkenyereye, G. P. Joshi, and E. Yang, " Mitigating and monitoring smart city using internet of things," *Computers, Materials & Continua*, vol. 65, no. 2, pp. 1059–1079, 2020.

[13] S. Milad, S. Taheri, and J. S. Yuan. "Utilizing transfer learning and homomorphic encryption in a privacy preserving and secure biometric recognition system," *Computers*, vol. 8, no. 1, 2019.

[14] S. Ahmad, S. Khan, M. F. AlAjmi, A. K. Dutta, L. Minh Dang *et al.*, "Deep learning enabled disease diagnosis for secure internet of medical things," *Computers, Materials & Continua*, vol. 73, no. 1, pp. 965–979, 2022. doi:10.32604/cmc.2022.025760

[15] Y. Yu, M. H. Au, G. Ateniese, X. Huang, W. Susilo *et al.*, "Identity-based remote data integrity checking with perfect data privacy preserving for cloud storage," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 4, pp. 767–778, April 2017.

[16] M. Du, Q. Wang, M. He, and J. Weng, "Privacy-preserving indexing and query processing for secure dynamic cloud storage," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 9, pp. 2320–2332, Sept. 2018.
[17] W. Shen, J. Yu, H. Xia, H. Zhang, X. Lu *et al*., "Light-weight and privacy-preserving secure cloud auditing scheme for group users via the third party medium," *Journal of Network and Computer Applications*, vol. 82, pp. 56–64, 2017.
[18] K. Liang, X. Huang, F. Guo, and J. K. Liu, "Privacy-preserving and regular language search over encrypted cloud data," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 10, pp. 2365–2376, 2016.
[19] W. Itani, A. Kayssi, and A. Chehab, "Wireless body sensor networks: Security, privacy, and energy efficiency in the era of cloud computing," *Cyber Law, Privacy, and Security: Concepts, Methodologies, Tools, and Applications*, pp. 731–763, 2019.
[20] J. R. Gudeme, S. K. Pasupuleti, and R. Kandukuri, "Public integrity auditing for shared data with efficient and secure user revocation in cloud computing," in *Proc. 2018 International Conference on Advances in Communication and Computing Technology (ICACCT)*, 2018, pp. 588–593.
[21] Y. Fan, X. Lin, G. Tan, Y. Zhang, W. Dong, and J. Lei, "One secure data integrity verification scheme for cloud storage," *Future Generation Computer Systems*, vol. 96, pp. 376–385, 2019.
[22] Y. Zhang, C. Xu, X. Liang, H. Li, Y. Mu, and X. Zhang, "Efficient public verification of data integrity for cloud storage systems from indistinguishability obfuscation," *IEEE Transactions on Information Forensics and Security*, vol. 12, pp. 676–688, 2016.
[23] M. Imran, H. Hlavacs, B. J. I. Ul-Haq, F. A. Khan, and A. Ahmad, "Provenance based data integrity checking and verification in cloud environments," *PLoS one*, vol. 12, 2017.
[24] M. K. Hasan, S. Islam, I. Memon, A .F. Ismail, S. Abdullah, A. K. Budati, and N. S. Nafi, "A novel resource oriented DMA framework for internet of medical things devices in 5G network," *IEEE Transactions on Industrial Informatics*, 2022.
[25] M. K. Hasan, M. Shafiq, S. Islam, B. Pandey, Y. A. B. El-Ebiary, N. S. Nafi, R. C. Rodriguez, and Vargas, "Lightweight cryptographic algorithms for guessing attack protection in complex internet of things applications," *Complexity*, 2021.
[26] S. Ahmad, H. A. Abdeljaber, J. Nazeer, M. Y. Uddin, V. Lingamuthu, and A. Kaur, "Issues of clinical identity verification for healthcare applications over mobile terminal platform," *Wireless Communications and Mobile Computing*, Apr. 2022.
[27] Y. F. Khan, B. Kaushik, M. K. I. Rahmani, and M. E. Ahmed, "Stacked deep dense neural network model to predict alzheimer's dementia using audio transcript data," *IEEE Access*, vol. 10, pp. 32750–32765, 2022. doi: 10.1109/ACCESS.2022.3161749
[28] A. Alharbi, K. Equbal, S. Ahmad, H. U. Rahman, and H. Alyami, "Human gait analysis and prediction using the levenberg-marquardt method," *Journal of Healthcare Engineering*, Feb. 2021.
[29] A. Bahaa, A. Sayed, L. Elfangary, and H. Fahmy, "A novel hybrid optimization enabled robust CNN algorithm for an IoT network intrusion detection approach," *PLos One*, vol. 17, no. 12, e0278493, 2022.