Applied Science & Engineering Journal for Advanced Research
ISSN (Online): 2583-2468
Volume-1 Issue-5 || September 2022 || PP. 1-14 DOI: 10.54741/asejar.1.5.1

Jaccard Index Cat Gradient Boosting Classification for Secured Big Data Communication

S L Swapna[1] and V Saravanan[2]
[1]*Research Scholar, Hindusthan College of Arts and Science (Autonomous), Coimbatore, India*
[2]*Professor, Hindusthan College of Arts and Science (Autonomous), Coimbatore, India*

[1]*Corresponding Author: swapnamartin2003@gmail.com*

Received: 01-08-2022	Revised: 22-08-2022	Accepted: 12-09-2022

ABSTRACT
Big data is observed as a novel field dealing with datasets that are too complex in providing indispensable services for daily chores and also discovering hidden patterns. Network security has become a major issue due to big data analytics, which offers unlimited research potential. More specifically, secure data communication without a third party is a major concern. Also, as large, heterogeneous, and complex data sets emerge, existing security mechanisms cannot provide or address network threats quickly or accurately. Therefore, along with the decrease in time, accuracy and error rate are other research concerns. Accordingly, an accurate and timely big data-based secure method called Jaccard Index Cat Gradient Boosting Classification-based Secured Data Communication (JICGBC-SDC) using the Internet of Things is presented. Firstly, for each cloud user, user registration is performed by acquiring information from various sensors. Second, information is collected from the registered cloud users by means of the Jaccard Index Cat Gradient Boosting Classifier algorithm. Such a proposed algorithm imposes a lower error rate and minimizes classification time, ensuring the most reliable and secured data communication between cloud users. To ensure secure data communication, weak learners' results are combined to form a strong classifier. The proposed method is implemented in Java and tested on the CloudSim simulator for classification accuracy, classification time, and error rate. The experimental results reveal the JICGBC-SDC method increases the performance of secured data communication for error rate by 77%, classification time by 79% and classification accuracy by 25% as compared to the state-of-the-art work.

Keywords*: big data, cloud, jaccard index, cat gradient boosting, data communication, internet of things*

I. INTRODUCTION

The Internet of Things (IoT) has a swift extension in the evolution of smart IoT-based physical objects are interconnected with sensors to gather, process, and store data. The collected data is further utilized for post-analysis and smarter decisions in the cloud computing environment. However, security poses a major threat to IoT-based sensors as far as healthcare data is concerned [1]. In addition, IoT-based sensors also connect with big data using the Internet for secure data communication between cloud users. Therefore, along with the minimization of time, accuracy, error rate, and security, there are other research concerns as well [2].

An energy-efficient and big data-based secure framework (EBDS) was introduced in [3] with the Internet of Things for the green environment. An IoT-based sensor was employed for data gathering, followed by data routing, by means of a Dijkstra-based optimal algorithm. With the aid of this algorithm, overhead and energy consumption involved in reliable transmission were said to be reduced. Despite improvements observed in both energy consumption and overhead, the accuracy level was not improved for the big data-based secure framework.

Big data stored on different types of heterogeneous network node types is prone to data security and privacy issues like virus infection. The users involved in this type of heterogeneous network were split into five states consistent with reactions to data viruses, namely vulnerable, infectious, doubtful, immune, and recoverable. Also, a novel model was introduced for learning about data virus propagation. Furthermore, an incentive-based protection and recovery strategy was implemented in [4] to deal with virus spread and thus ensure data security. Finally, a Protection and Recovery Strategy (PRS) was also introduced to reduce the number of infected users. However, the time complexity was not reduced by the incentive-based protection and recovery strategy.

Bigdata systems were employed to store, manage, and use data from large-scale WSNs. In big data systems, the access control technology affects the system's performance [5]. The data processing flow of the access control approach in bigdata systems examined the time complexity and affected the system performance. The big data security access control algorithm was introduced in [6] and depends on memory index acceleration. Though the time complexity was reduced, the security level was not improved by the big data security access control algorithm.

To address big data security issues in the Cloud, the Secure Authentication and Data Sharing in the Cloud (SADS-Cloud) architecture was introduced in [7].The designed architecture consisted of three processes, namely big data outsourcing, big data sharing, and big data management. In big data outsourcing, the data owners were registered with the Trust Center using the SHA-3 hashing algorithm. The MapReduce model was employed to divide the input file into fixed-size data blocks. In big data sharing, data users participate in secure file retrieval. However, the computational complexity was not reduced by the SADS-Cloud architecture.

To address the issues raised in the preceding reviews, a JICGBC-SDC method is presented, along with the novel contributions listed below.

- To improve secure big data communication between users, a JICGBC-SDC method is designed on the basis of three distinct processes: user registration, data collection, and data communication.
- Jaccard Index A Cat Gradient Boosting Classifier is applied for secure big health data communication. The Cat Gradient Boosting Classifier algorithm initially uses a Jaccard Similarity Index that analyzes the receiver ID and the registered ID between the testing and training sample. By minimizing the loss function, Cat Gradient Boosting combines poor classification performance with strong results. This helps to improve the classification accuracy involved during the communication process.
- An extensive experiment was conducted to estimate the performance of the JICGBC-SDC method and other related work. The obtained result shows that our proposed JICGBC-SDC method provides better performance in terms of classification accuracy, classification time, and error rate.

The paper is organized into five different sections. Related works are reviewed in Section II. An elaborate description of the proposed JICGBC-SDC method with the diagram representation is provided in Section III. The experimental evaluation is summarized in Section IV, followed by the results being discussed in Section V. Finally, the paper is concluded in Section VI.

II. RELATED WORKS

With the swift deployment of science and technology, the evolution of mobile networks is found to be more accelerated than ever before. To be more specific, 5G networks have been specifically utilized in several first-tier cities, ensuring communication between people in a more efficient manner. However, owing to the fact that huge wireless terminal devices are associated with the communication network, the resources of the wireless spectrum are found to be very scarce.

An introduction to safe and secure transportation strategies in an intelligent manner employing Deep Learning Models was proposed in [8]. An integration of deep neural networks and deep reinforcement learning based on a loss function was presented in [9], ensuring a higher success rate and faster convergence speed. Despite the improvement observed in convergence speed, the quality of service was not provided. A survey on the use of machine learning to ensure connectivity and high quality of service (QoS) was investigated in [10].

The extent of data produced and dispensed by businesses, public institutions, several profit and non-profit organizations, public sectors, and private sectors has grown exponentially. This data comprises content involving text, video, audio, and multimedia content on a range of manifestos. In [11], a cutting-edge review of big data challenges and big data analytics for secure communication was created.

With the enormous broadening of data, secure data communications are gaining progressively more significance in examining information resources and aiding users to accomplish their daily chores in an efficient manner. A framework was proposed in [12] with the objective of applying numerous unsupervised learning techniques, consisting of both topic modeling and log mining. Moreover, the application of the results to a greater number of user session data infers that these techniques would be found to be more appropriate in making secure searching and task recommendations, attaining a notable enhancement over a strong baseline.

A survey of implementations of IoT and AI for remote healthcare data communication was presented in [13]. The contemporary amalgamation of numerous services has resulted in big data sets consisting of both individuals and organizations. This epidemic widening in the size of data has validated businesses' acquisition of inflated comprehension of large data-sets. This enormous advancement has engendered a great number of issues, specifically when taking into account storage, security analysis, and privacy preservation.

In [14], a Dynamic Maximum Transmission Unit (DMTU) scheme was proposed that in turn handled packet drops in IPv6 networks by means of adjusting the maximum transmission unit in a dynamic manner. This was performed based on the

size of the incoming data packet and, as a result, the packet drop rate was found to be significantly less. Also, optimized data communication was also ensured by means of validation.

In [15], a novel homomorphic encryption model was discovered to be particularly user-friendly for the Paillier crypto-system. This method was called "Secure Analytic Using Vector-based Homomorphic Operation (SAVHO) and was found to be highly resistant against numerous attack types.

The administration of the epidemic heightening of data that Next Generation Sequencing algorithms generate has become a great issue for researchers who are compelled to go through an ocean of complicated data with the purpose of acquiring new perceptions to resolve the confidence of human diseases. In this aspect, accurate identification of genome data management is a prerequisite to bestowing efficient big data-based solutions [16]. Robust information systems that utilize big data techniques were provided for security purposes.

The fourth industrial revolution comprises of the advancements made in the manufacturing sectors employing cyber-physical systems with the purpose of enhancing manufacturing potentialities and pliability to modify production swiftly and effectively with respect to transposing conditions and insistences. Also, cyber security risks [17] were analyzed with the probable solutions to safeguard against the different types of attacks and enhance the robustness of the overall system employing machine learning. Several threats and countermeasures were addressed in [18] in terms of security and privacy using machine learning techniques.

In the purview of the COVID-19 epidemic, interference of upcoming techniques has been heavily increased, and post-pandemic, a large shift in technology is anticipated in bestowing both information and communication between healthcare personnel. With this type of heavily dynamic and heterogeneous type of structure and swift transformation in digital structure, providing security resource-restrictions and performance implications are said to be major issues to be handled hand in hand.

In [19], prevailing security, privacy, and authentication mechanisms, their strengths and weaknesses with respect to IoT or IoMT, and big data were analyzed first. Second, on the basis of the drawbacks, potential solutions were also provided by employing machine learning. An elaborate and comprehensive survey of machine learning techniques in the context of security and privacy was investigated in [20].

Yet another holistic survey concerning security threats and countermeasures provided by means of AI techniques was elaborated in detail in [21]. An in-depth experimental analysis for providing cyber security via federated deep learning was proposed in [22]. A survey of secure federated learning mechanisms for secure data communication was investigated in [23].

Motivated by the preceding works, a method called Jaccard Index Cat Gradient Boosting Classification (JICGBC-SDC) using IoT is proposed to ensure accurate and timely classification of authorized and unauthorized users using machine learning. The elaborate description of the method with the diagram and algorithm is explained in detail in the coming sections.

III. METHODOLOGY

Modern healthcare devices possess the potential to acquire medical data involving blood pressure, heart beat rate, glucose, etc. by using sensors. This data is then said to be communication between users for further processing and analysis. However, with the volume and size of data increasing, secure data communication performed using different techniques provided in the literature is not found to be robust enough to different types of attacks.

To address the secure data communication aspect, a method called Jaccard Index Cat Gradient Boosting Classification based Secured Data Communication (JICGBC-SDC) is proposed. The JICGBC-SDC method consists of three phases. They are user registration, data collection, and data communication. The elaborate description of the JICGBC-SDC method is discussed in detail in the following sections, followed by a system model.

1.1 Cloud-Enabled IOT System Model

The cloud-enabled structural design consists of distinct entities contributing the secured big data communication. The system model comprises of different numbers of cloud user's '$CU = CU_1, CU_2, \ldots, CU_n$' who wants to store big data '$D = D_1, D_2, \ldots, D_n$' in cloud server '$CS$' in a secured manner. The cloud service provider 'CS' on the other hand provides various services like, user registered details (i.e., data) 'D' stored in the cloud server 'CS', performing classification and finally secured data communications between cloud users. Figure 1 given below shows the Cloud-enabled IoT System Model.

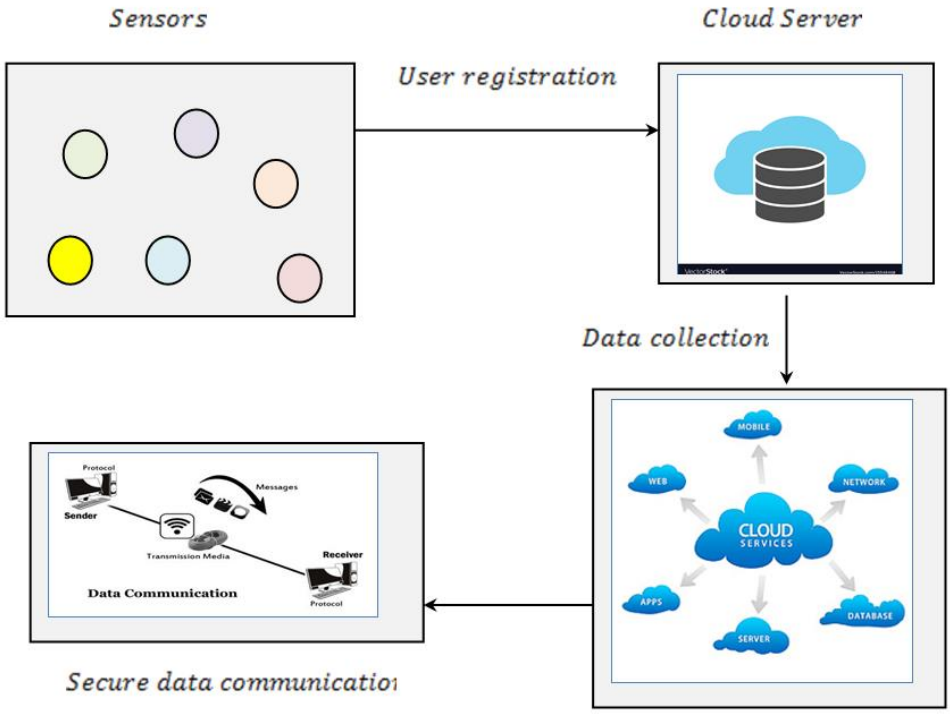

Figure 1: Cloud-enabled IoT System Model

As shown in the above Cloud-enabled IoT System Model employed in our work, first, to start with, the data from the respective cloud users sensors (i.e., from 14 sensors as provided in table 1) are acquired as input and stored in the cloud server. Followed by which, second the actual data collection is performed using boosting classification model. Finally, secure data communication is established. The elaborate description of the proposed method is discussed in detail in the following sections.

1.2 User Registration

In the JICGBC-SDC method, initially, the cloud user acquires the data from the corresponding sensors and enters the details into the cloud server (CS). Figure 2 given below shows the structure of user registration followed in our proposed JICGBC-SDC method.

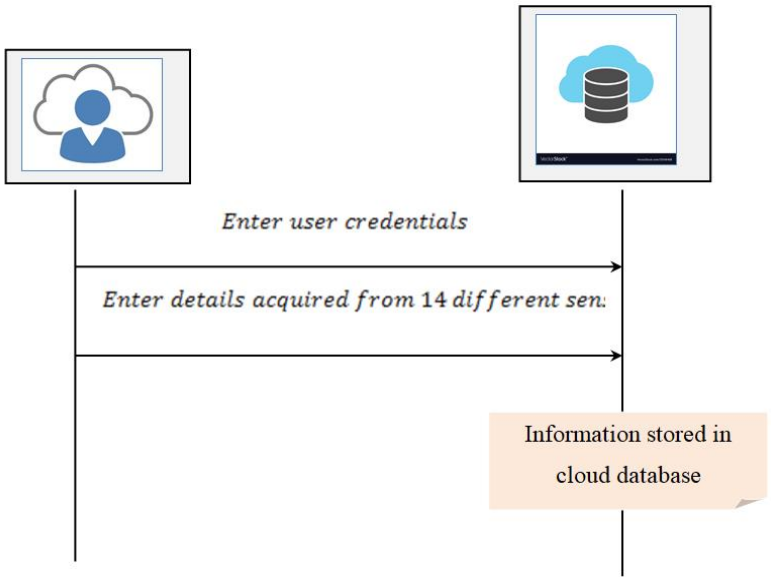

Figure 2: Structure of JICGBC-SDC user registration

As shown in the above figure, two types of data based on acceleration and gyroscope data are taken for the classification of given activities. Initially, the user's credentials like ID, password, secure ID, current timestamp, and email ID along with the fourteen sensor details provided in table 1 are registered in the registration phase. These user's credentials obtained from the corresponding sensors (i.e., 'S_{alx}', 'S_{aly}', 'S_{alz}',…. '$S_{subject}$') data are stored in the cloud server for future reference.

Table 1: Dataset description

S. No	Features or parameters	Description
1	alx	Acceleration from left-ankle sensor (X axis)
2	aly	Acceleration from left-ankle sensor (Y axis)
3	alz	Acceleration from left-ankle sensor (Z axis)
4	glx	Gyro from the left-ankle sensor (X axis)
5	gly	Gyro from the left-ankle sensor (Y axis)
6	glz	Gyro from the left-ankle sensor (Z axis)
7	arx	Acceleration from right-ankle sensor (X axis)
8	ary	Acceleration from right-ankle sensor (Y axis)
9	arz	Acceleration from right-ankle sensor (Z axis)
10	grx	Gyro from the right-ankle sensor (X axis)
11	gry	Gyro from the right-ankle sensor (Y axis)
12	grz	Gyro from the right-ankle sensor (Z axis)
13	Activity	Corresponding activity
14	Subject	Volunteer subjects (1 – 9)

As given in the above table, the collected dataset consists of recordings of body motion and vital signs acquired for ten distinct volunteers of different profiles while performing twelve different types of physical activities (i.e., standing still, sitting and relaxing, lying down, walking, climbing stairs, waist bends forward, frontal elevation of arms, knees bending, cycling, jogging, running, jumping front and back, respectively). Wearable sensors based on shimmer2 [BUR10] were utilized for recordings. To be more specific, the sensors were positioned on the subject's chest, right wrist, and left ankle. These different types of sensors permit us to acquire the distinct motion recorded by different body parts, namely, the acceleration and rate of turn, hence acquiring the body dynamics in a better manner. All the above sensors, comprising different types of modalities, were recorded at a sampling rate of 50 Hz, and also, each session was recorded with the assistance of a video camera.

1.3 Jaccard Index Cat Gradient Boosting Classifier-Based Secure Data Communication

In the second phase, the data is collected from the registered user. The second process of the proposed JICGBC-SDC method is to perform actual data collection based on the registered user details. Here, the Jaccard Index Cat Gradient Boosting Classifier is utilized to perform strong classification by combining weak learners for secured data communication. In our work, the Jaccard similarity index is considered the weak learner that analyzes the receiver ID with the registered ID to classify the user as either an authorized user or an unauthorized user. This Jaccard similarity index is formulated as given below.

$$J(RID, RegID) = \frac{|RID \cap RegID|}{|RID \cup RegID|} \quad (1)$$

$$JSI(RID, RegID) = \frac{|RID \cap RegID|}{|RID| + |RegID| - |RID \cap RegID|} \quad (2)$$

From the above equations (1) and (2), the Jaccard Similarity Index 'JSI' is obtained based on two distinct sample sets i.e., receiver ID 'RID' with the registered ID '$RegID$'. With the index value ranging from '$0\ to\ 1$', the close to index value result being '1', more similar are the two sets of big data i.e., more similar are the receiver ID with registered ID and vice versa. Figure 3 given below shows the structure of ensemble classifiers used for Jaccard Index Cat Gradient Boosting Classifier-based Secure Data Communication.

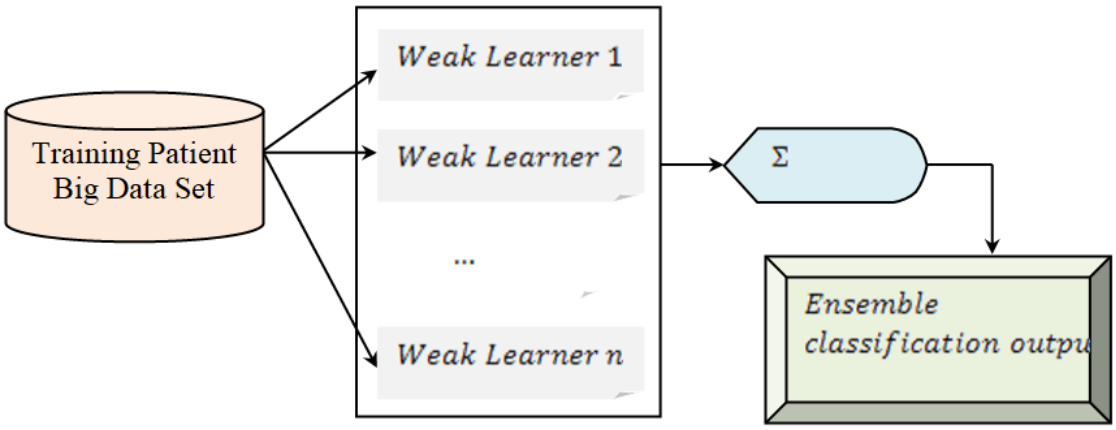

Figure 3: Structure of Ensemble Classifiers

Figure 3 given above shows the structure of ensemble classification to ensure secure data communication for a set of training samples i.e. a number of patient big data '$D = D_1, D_2, \ldots, D_n$' acquired from distinct cloud user's '$CU = CU_1, CU_2, \ldots, CU_n$'. The weak learner is a base classifier that lacks providing accurate classification results. On the contrary, a boosting is an ensemble classifier that provides accurate classification results by combining a set of weak classifiers.

Cat Gradient Boosting is an ensemble classifier that converts weak learners into strong learns. Therefore, the proposed method uses the Cat Gradient boosting ensemble algorithm to improve the performance of classification. The ensemble algorithm in this work utilizes the set of weak learners '$WL = WL_1, WL_2, \ldots, WL_n$' to train the number of patient big data '$D = D_1, D_2, \ldots, D_n$' and combined into a strong one to ensure secure data communication between cloud users. The Cat Gradient Boosting uses weak learners as jaccard similarity index for classifying training samples. Figure 4 given below shows the block diagram of Jaccard Index Cat Gradient Boosting Classifier model.

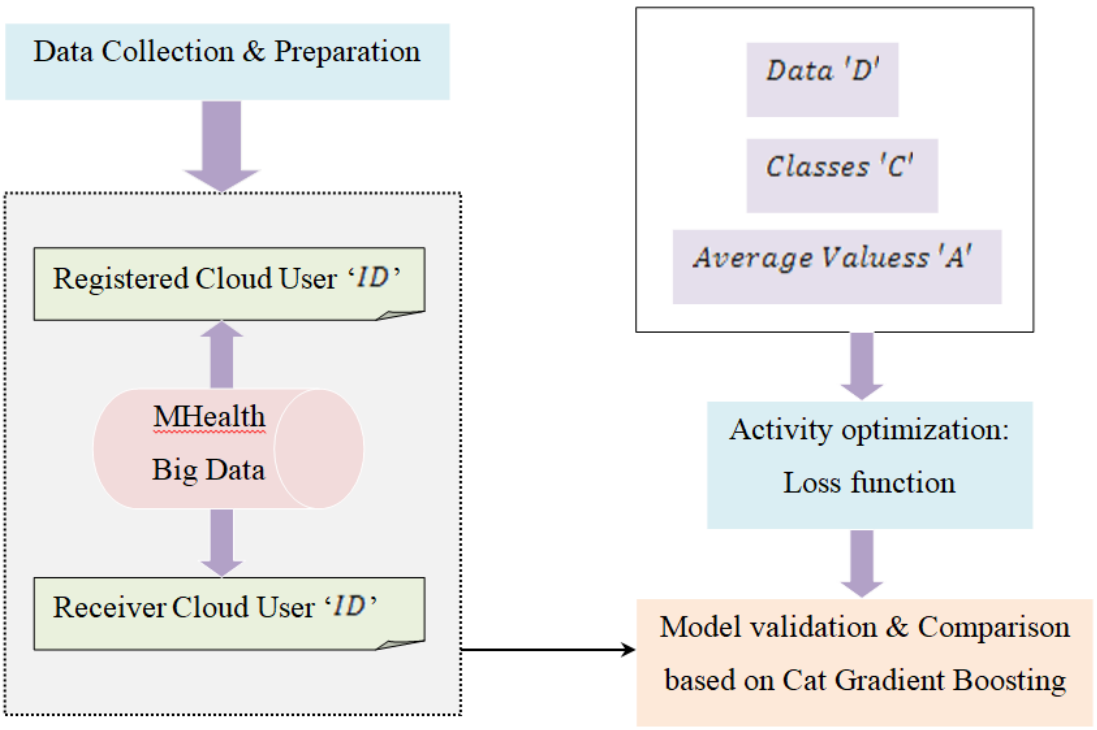

Figure 4: Jaccard Index Cat Gradient Boosting Classifier

As shown in the above figure, the jaccard similarity index defines the number of classes '$C = C_1, C_2, \dots, C_n$' and their average values '$A = A_1, A_2, \dots, A_n$'. Followed by, the training big data samples are analyzed with the testing big data value and obtain the final communication outcomes using CatBoost function.

$$Res = argmax\left[a_l^i[(D_i)|(D_i)]\right] \quad (3)$$

Let us consider with a set '$\{a_i, b_i\}$' of input values 'a_i' and expected output values '$b_i, i \in \{1,2,..,n\}$'. The Gradient boosting in Cat function then iterative construct a multitude of activities '$A^1, A^2, \dots, A^n$', given a loss function '$\mathcal{L}\ (b_i, A^t)$'. Here, the loss function possesses two distinct values '$i-th$' expected output value 'b_i' and the '$t-th$' activity that evaluates 'b_i'. Then, with the assumption that activity 'A^t' can enhance the secure data communication between cloud users (i.e., between patients, between patient and doctors), of 'b_i', let us find another activity as given below.

$$A^{t+1} = A^t + H^{t+1}(a) \quad (4)$$

Let us further formulate the above activity with the objective of minimizing the loss function and formulated as given below.

$$H^{t+1} = argmin\ E\mathcal{L}(b, A^t) \quad (5)$$

From the above equation (5), 'H' refers to the candidate set Decision Trees we are estimating to select one cloud users big data or information to add to the ensemble. Finally, in the third phase, the actual data communication is performed where the results of weak learners are combined to form strong classifier with higher accuracy and minimum error rate. This in turn helps to ensure secured data communication.

To ensure secure data communication, CatBoost encodes the values of categorical variables via an indicator function '$\mathbf{1}$'. Here, the indicator function '$\mathbf{1}_{a=b}$' refers to a function of one variable '$\mathbf{1}$' that possess the value '$\mathbf{1}$' when '$a = b$' and '$\mathbf{0}$' otherwise. The indicator function in the Cat Gradient Boosting for secure data communication plays a vital role for efficient mapping between the cloud users' data for categorical representation.

To be more specific the indicator function '$I_{a_j^i = a_l^i}$', that takes the value '$\mathbf{1}$' when the '$i-th$' strong learner of CatBoost's cloud users input vector 'a_j' is equal to the '$i-th$' strong learner of the cloud users input vector 'a_l'. Then, with the above assumption, with the training big data 'D' and the indicator function '$I_{a_j^i = a_l^i}$', the mathematical formulation for the encoded value for secure data communication is stated as given below.

$$a_l^i = \frac{\sum_{a_j \epsilon\, D_l} I_{a_j^i = a_l^i} \cdot b_j + paramC}{\sum_{a_j \epsilon\, D_l} I_{a_j^i = a_l^i} + param} \quad (6)$$

From the above equation (6), 'C' defines the number of prior classes '$C = C_1, C_2, \dots, C_n$' commonly set to the average values '$A = A_1, A_2, \dots, A_n$' of instances in the dataset 'DS' and '$param$' as a parameter greater than '$\mathbf{0}$'. Finally, the encoding categorical representation possesses certain property as given below.

$$E\left(a^i | b = RID\right) = E\left(a_l^i | b_l = RegID\right) \quad (7)$$

The CatBoost's encoding function as given above (7) satisfies this property (i.e., analyzing receiver ID 'RID' with the registered ID '$RegID$' to ensure secure data communication between cloud users. The pseudo code representation of Jaccard Index Cat Gradient Boosting for secured data communication in cloud environment is given below.

INPUT: DATASET 'DS', SENSORS '$S = S_1, S_2, \ldots, S_n$', '$n = 14$', CLOUD USER'S '$CU = CU_1, CU_2, \ldots, CU_n$', BIG DATA '$D = D_1, D_2, \ldots, D_n$', CLOUD SERVER '$CS$'

OUTPUT: ACCURATE AND TIMELY SECURED DATA COMMUNICATION

STEP 1: **INITIALIZE** RECEIVER ID 'RID', REGISTERED ID '$RedID$'
STEP 2: **BEGIN**
STEP 3: **FOR** EACH DATASET 'DS' WITH SENSORS 'S'
STEP 4: **FOR** EACH RECEIVER ID 'RID' AND REGISTERED ID '$RedID$'

//USER REGISTRATION
STEP 5: STORE SENSOR DETAILS 'S_{alx}', 'S_{aly}', 'S_{alz}',…. '$S_{subject}$' IN THE CLOUD SERVER

//DATA COLLECTION
STEP 6: EVALUATE JACCARD FUNCTION AS IN EQUATION (1)
STEP 7: EVALUATE JACCARD SIMILARITY INDEX AS IN EQUATION (2)
STEP 8: FORMULATE CATBOOST FUNCTION AS IN EQUATION (3)
STEP 9: EVALUATE ACTIVITY FORMULATION AS IN EQUATION (4)

//DATA COMMUNICATION
STEP 10: EVALUATE ACTIVITY FORMULATION WITH MINIMUM LOSS FUNCTION AS IN EQUATION (5)
STEP 11: FORMULATE CATBOOST'S ENCODING FUNCTION AS IN EQUATION (6)
STEP 12: **IF** '($b[RID = RegID]$)'
STEP 13: **THEN** CLOUD USER IS AUTHORIZED USER
STEP 14: PERFORM SECURE DATA COMMUNICATION
STEP 15: **END IF**
STEP 16: **IF** '($b[RID \neq RegID]$)'
STEP 17: **THEN** CLOUD USER IS NOT AN AUTHORIZED USER
STEP 18: NO DATA COMMUNICATION BETWEEN CLOUD USERS
SEP 19: **END IF**
STEP 20: **END FOR**
STEP 21: **END FOR**
STEP 22: **END**

Algorithm 1: Jaccard Index Cat Gradient Boosting Classification

As given in the above algorithm with the objective of ensuring accurate and timely communication of healthcare big data between cloud users in cloud environment, Jaccard Index Cat Gradient Boosting Classifier is applied to the input healthcare data. As healthcare data are said to be highly prone to be intercepted by the adversaries during the communication between users, these data has to be secured by means of security mechanisms. In this work, Jaccard Index Cat Gradient Boosting is used where with the aid of the Cat Gradient Boosting strong learners are ensemble by means of Jaccard Similarity Index function (via means of weak learners). Here, the weak learners obtained via Jaccard Similarity Index function is ensemble via Cat Gradient Boosting therefore ensuring secure data communication.

IV. RESULTS AND DISCUSSION

In this section, the experimental evaluation of the proposed Jaccard Index Cat Gradient Boosting Classification based Secured Data Communication (JICGBC-SDC) method and the existing Energy-efficient Big Data-based Secure (EBDS) [1] method, Incentive-based Protection and Recovery Strategy (Incentive-based PRS) [2] is implemented in the Java language via CloudSim simulation. In order to conduct the experiment and perform secure big data communication between the cloud users, the Mobile Health dataset is used.

The dataset consists of 14 attributes and 12,15,745 instances. The main aim of the dataset is to record several physical activities for ten volunteers of diverse profiles. The dataset details are provided in table 1. Experimental evaluation is carried out on factors such as classification accuracy, classification time, and error rate with respect to the number of users.

The quantitative performance evaluation of the JICGBC-SDC method and the existing EBDS [1] and incentive-based PRS [2] are compared with certain parameters such as classification accuracy, classification time, and error rate with respect to

the number of users. The performance of proposed and existing methods is discussed with the help of a table and graphical representation.

a. Case Scenario 1: Classification Accuracy

Initially, to analyze the proposed JICGBC-SDC method and make a comparison with two other existing methods for secure data communication in the cloud, users have to be classified as either authorized or unauthorized. Only after the successful classification, the actual data communication has to be done. Therefore, the accuracy rate with which the classification is ensured plays a major role during secure data communication. This classification accuracy is mathematically expressed as given below.

$$CA = \sum_{i=1}^{n} \frac{CU_{AC}}{CU_i} \quad (8)$$

From the above equation (8), the classification accuracy 'CA' is measured based on the cloud users involved in the simulation process 'CU_i' and the cloud users accurately classified 'CU_{AC}'. It is measured in terms of percentage (%). This experiment compares our proposed JICGBC-SDC method with the existing EBDS [1] and Incentive-based PRS secure data communication based methods given in the literature. The comparison is based on classification accuracy. Table 2 presents the result comparison of our method with other previous secured data communication methods.

Table 2: Comparison between various secured data communication methods based on classification accuracy

Cloud Users	Classification accuracy (%)		
	JICGBC-SDC	EBDS	Incentive-based PRS
10000	98.45	97.32	96.53
20000	97.15	91.25	85.15
30000	96.85	90.15	84.45
40000	96.00	90.00	84.00
50000	95.55	89.15	82.15
60000	95.15	88.00	81.00
70000	94.35	86.45	79.15
80000	94.00	85.15	77.00
90000	93.55	84.00	75.25
100000	93.00	82.25	72.00

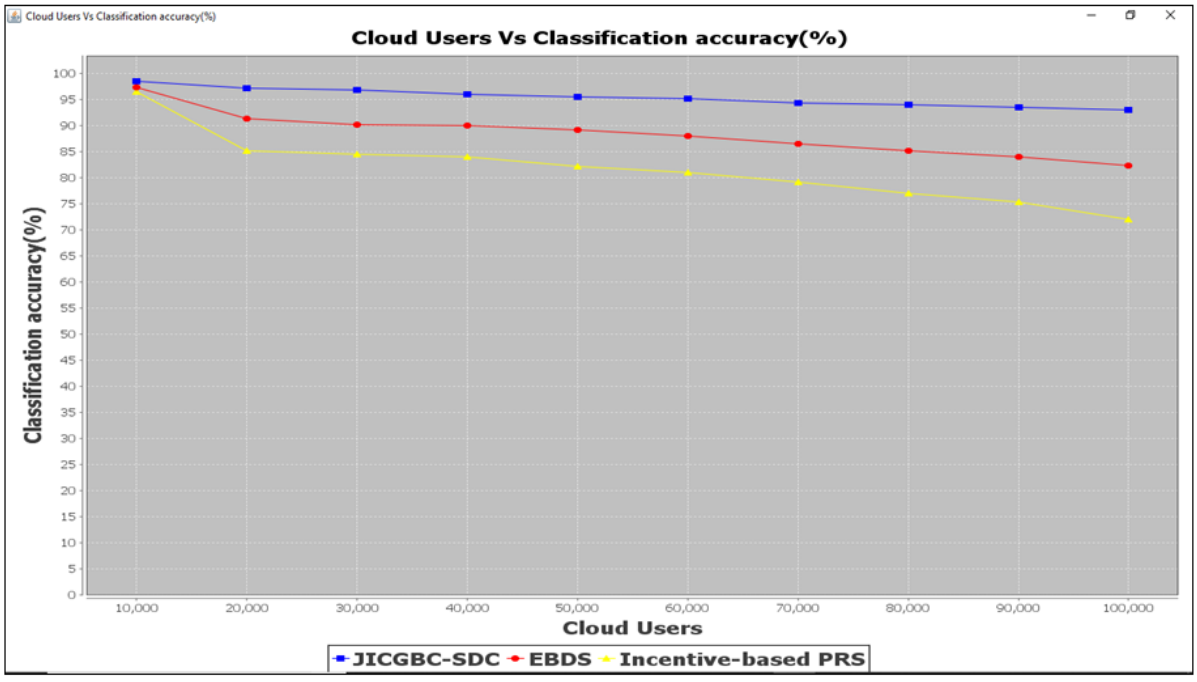

Figure 5: Classification accuracy and varying cloud users

Figure 5, given above, shows the classification accuracy for varying numbers of cloud users in the range of 10,000 to 100,000. From the figure, it is inferred that increasing the frequency of cloud users results in a small decrease in the classification accuracy performance. However, with simulations conducted for 10,000 users, 98455 cloud users were properly found to be authorized as authorized using JICGBC-SDC, 97325 cloud users were properly found to be authorized as authorized using [1], and 96535 cloud users were properly found to be authorized as authorized using [2]. With this, the accuracy rate of classification using the three methods, JICGBC-SDC, [1] and [2], was observed to be 98.455, 97.32%, and 96.53%, respectively. With this, the classification accuracy using JICGBC-SDC was found to be better than [1] and [2]. The reason behind the improvement was the application of the Cat Gradient Boosting classifier that first obtained the weak class learners using the Jaccard Similarity Index. Strong learners were formed by utilizing Cat Gradient Boosting, which in turn increased the classification accuracy using JICGBC-SDC by 8% compared to [1] and 17% compared to [2] respectively.

b. Case Scenario 2: Classification Time

Second, the more swift the classification between authorized and unauthorized cloud users, the better the method is said to be in ensuring secure data communication between users. To analyze this swift classification being made between authorized and unauthorized users, classification time is required. The classification time here refers to the time consumed in classifying between the authorized and unauthorized users waiting for communication. This is mathematically formulated as given below.

$$CT = \sum_{i=1}^{n} CU_i * Time\ [classification] \quad (9)$$

From the above equation (9), the classification time 'CT' is measured based on the cloud users involved in the simulation process 'CU_i' and the time consumed in classification '$Time\ [classification]$'. It is measured in terms of milliseconds (ms). This experiment compares our proposed JICGBC-SDC method with the existing EBDS [1] and Incentive-based PRS secure data communication methods given in the literature. The comparison is based on classification time. Table 3 presents the result of the comparison of our method with other previous secure data communication methods.

Table 3: Comparison between various data secure communication methods based on classification time

Cloud Users	Classification time (ms)		
	JICGBC-SDC	EBDS	Incentive-based PRS
10000	5000	8000	11000
20000	5135	8135	11835
30000	6245	8895	12145
40000	7135	10235	13835
50000	8245	12455	15625
60000	10355	14315	18355
70000	11215	16255	21435
80000	13235	18155	23555
90000	14155	19325	26245
100000	15215	20145	29145

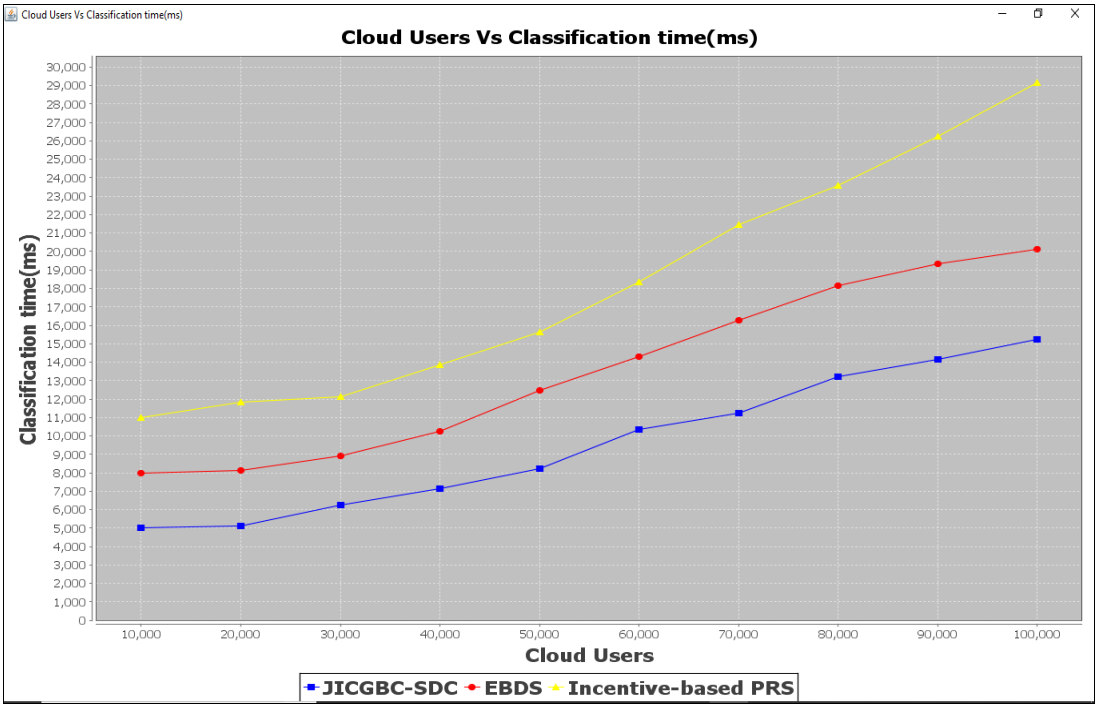

Figure 6: Classification time and varying cloud users

Figure 6 given above illustrates the time consumed in classifying the cloud users between authorized and unauthorized users while performing secured data communication. To conduct the simulation, 100,000 distinct cloud users were employed for data communication. From the figure, the classification time is found to be directly proportional to the number of cloud users. In other words, increasing the frequency of cloud users causes an increase in the number and size of data to be transmitted, which in turn causes an increase in the classification time also. However, with simulations performed with a sample of 10,000 cloud users the time consumed in classifying between authorization and non-authorization for single cloud user was found to be '$0.05ms$' using JICGBC-SDC, '$0.08ms$' using [1] and '$0.11ms$' using [2] respectively. The overall classification time for 10000 cloud users into authorized and non-authorized was observed to be '$5000ms$', '$8000ms$' and '$11000ms$' respectively. From this results it is inferred that the classification time involved in the secure data communication process is comparatively less using JICGBC-SDC than [1] and [2]. The reason behind the minimization of the time involved in the classification for the data communication process was first introducing the user registration process. Only the registered cloud users were then involved in the data communication process. Also, the learners are boosted with no separate preprocessing mechanisms required owing to the application of the Cat Gradient Boosting classifier in the JICGBC-SDC method. With this, the time involved in classification processing using the JICGBC-SDC method is said to be reduced by 31% compared to [1] and 48% compared to [2] respectively.

c. Case Scenario 3: Error Rate

Finally, the error involved during secured big data communication is measured. This is because of the reason that there is said to be a small amount of wrong classification between authorized and unauthorized users, i.e., authorized users are classified as unauthorized users and vice versa, resulting in an error. Hence, the error rate has to be analyzed to estimate the efficiency of the method. This is mathematically stated as given below.

$$ER = \sum_{i=1}^{n} \frac{CU_{WC}}{CU_i} \quad (10)$$

From the above equation (10), the error rate 'ER' is measured based on the cloud users involved in the simulation process 'CU_i' and the cloud users wrongly classified 'CU_{WC}'. It is measured in terms of percentage (%). Finally, the experiment compares the proposed JICGBC-SDC method with the existing EBDS [1] and Incentive-based PRS secure data communication methods given in the literature. The comparison is based on the error rate. Table 4 presents the result of the comparison of our method with other previous secure data communication methods.

Table 4: Comparison between various data secure communication methods based on error rate

Cloud Users	Error rate (%)		
	JICGBC-SDC	EBDS	Incentive-based PRS
10000	2.35	3.85	6.35
20000	3.85	5.35	7.95
30000	5.35	7.25	10
40000	6.15	8	11.45
50000	7	8.85	13
60000	7.35	10.25	14.15
70000	8.55	11.55	16
80000	9.25	13	17.35
90000	10.45	14.55	19
100000	11	16	20.45

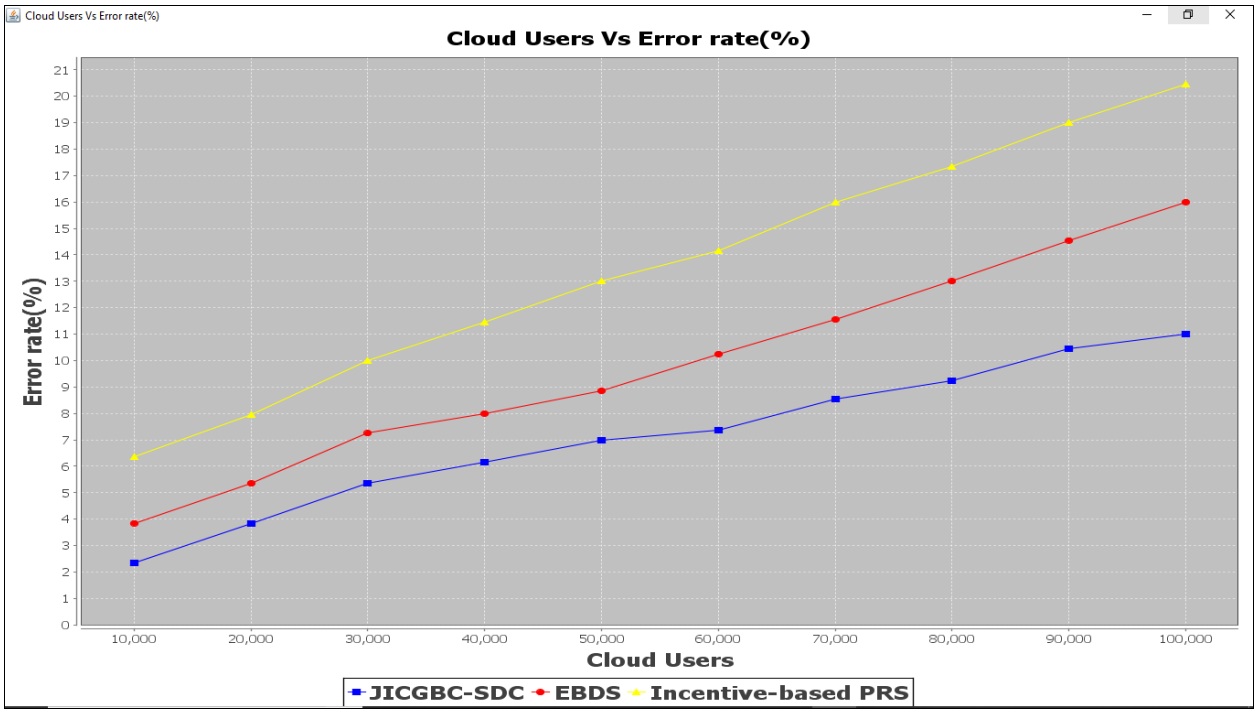

Figure 7: Error rate and varying cloud users

Finally, figure 7 given above shows the error rate involved during the classification process while performing secured data communication. While performing data communication between cloud users, a certain number of them may be wrongly classified as authorized users into un-authorized users or un-authorized users into authorized users, and accordingly the data communication process will proceed. From the figure, the error rate is found to be increasingly proportional to the number of cloud users. In other words, increasing the number of cloud users for simulation purposes causes a large number of them in the classification process and hence significant error rates occur with the testing data. However, simulations performed with 10,000 cloud users '2355' cloud users were wrongly classified as authorized to be un-authorized using JICGBC-SDC, '3855' cloud users were wrongly classified as authorized to be un-authorized using [1] and '6355' cloud users were wrongly classified as authorized to be un-authorized using JICGBC-SDC using [2] respectively. With this the error rate using the three methods were found to be '2.35%', '3.85%' and '6.35%' respectively. From this result, the error rate using JICGBC-SDC was said to be reduced upon comparison with [1] and [2]. The reason behind the minimum error rate using JICGBC-SDC was the application of the Jaccard Index Cat Gradient Boosting algorithm for secured data communication. Here, initially, Jaccard Index Cat Gradient Boosting was employed with the objective of turning ensemble weak learners into strong learners using Cat

Gradient Boosting via the Jaccard Similarity Index function. Moreover, the weak learner results obtained by means of the Jaccard Similarity Index function were then ensemble by utilizing Cat Gradient Boosting, thereby reducing the error rate and ensuring secure data communication between cloud users.

V. CONCLUSION

Machine learning and big data have been integrated in recent years to facilitate secure data communication and increase network performance. However, the increased nature and size of data poses research challenges to communities for accurate and timely secure data communication. This paper presents a Jaccard Index Cat Gradient Boosting Classification based Secured Data Communication (JICGBC-SDC) using the Internet of Things, which aims to reduce energy classification time and error rate with security for big data. The JICGBC-SDC method makes use of the Jaccard Index and improves the classification accuracy using the Jaccard Index Cat Gradient Boosting algorithm, which decreases the error rate during communication between cloud users. Moreover, the incorporation of Cat Gradient Boosting that performs mapping between the cloud users' data for categorical representation converts weak learners into strong learners and improves the data delivery performance with the least classification time. The simulation-based performance is evaluated, and the results indicate that the JICGBC-SDC method outperforms existing work for various network metrics.

REFERENCES

1. R John Martin. (2022). IoMT supported COVID care – Technologies and challenges. *International Journal of Engineering and Management Research, 12*(1), 125–131.
2. S L Swapna, & V Saravanan. (2021). Big data challenges and learning paradigms: A review. *Journal of University of Shanghai for Science and Technology, 23*, 36-45. doi: 10.51201/JUSST/21/11932.
3. Khalid Haseeb, Soojeong Lee, & Gwanggil Jeon. (2020). EBDS: An energy-efficient big data-based secure framework using Internet of Things for green environment. *Environmental Technology & Innovation, Elsevier, 20*, 1-19.
4. Youke Wu, Haiyang Huang, Ningyun Wu, Yue Wang, Md Zakirul Alam Bhuiyan, & Tian Wang. (2020). An incentive-based protection and recovery strategy for secure big data in social networks. *Information Sciences, Elsevier, 508*, 79-91.
5. S L Swapna, & V Saravanan. (2022). Survival analysis on secured data communication in cloud. *International Journal of Computer Applications, 183*(46), 31-35.
6. Jianhua Peng, Hui Zhou, Qingjie Meng, & Jingli Yang. (2020). Big data security access control algorithm based on memory index acceleration in WSNs. *EURASIP Journal on Wireless Communications and Networking, Springer, 2020*(90), 1-17.
7. Uma Narayanan, Varghese Paul, & Shelbi Joseph. (2020). A novel system architecture for secure authentication and data sharing in cloud enabled big data environment. *Journal of King Saud University - Computer and Information Sciences*, 1-15
8. A Jolfaei. (2021). Introduction to the special issue on deep learning models for safe and secure intelligent transportation systems. *IEEE Transactions on Intelligent Transportation Systems, 22*(7).
9. Ming LiID, & Hui Li. (2020). Application of deep neural network and deep reinforcement learning in wireless communication. *PLOS ONE*. Available at: https://doi.org/10.1371/journal.pone.0235447.
10. Saeed H. Alsamhi, Faris A. Almalki, Hatem Al-Dois, Soufiene Ben Othman, Jahan Hassan, Ammar Hawbani, Radyah Sahal, Brian Lee, & Hager Saleh. (2021). Machine learning for smart environments in B5G networks: Connectivity and QoS. *Computational Intelligence and Neuroscience, Hindawi*.
11. Uthayasankar Sivarajah, Muhammad Mustafa Kamal, Zahir Irani, & Vishanth Weerakkody. (2017). Critical analysis of big data challenges and analytical methods. *Journal of Business Research, Elsevier*.
12. Mohammad Alsulmi, & Reham Alshamarani. (2021). Framework for tasks suggestion on web search based on unsupervised learning techniques. *Journal of King Saud University – Computer and Information Sciences, Elsevier*.
13. Mazin Alshamrani. (2021). IoT and artificial intelligence implementations for remote healthcare monitoring systems: A survey. *Journal of King Saud University – Computer and Information Sciences, Elsevier*.
14. Ishfaq Hussain, & Janibul Bashir. (2021). Dynamic MTU: A smaller path MTU size technique to reduce packet drops in IPv6. *Journal of King Saud University – Computer and Information Sciences, Elsevier*.
15. Christiana Zaraket, Khalil Hariss, Maroun Chamoun, & Tony Nicolas. (2021). Cloud based private data analytic using secure computation over encrypted data. *Journal of King Saud University – Computer and Information Sciences, Elsevier*.

16. AnaLeón, & ÓscarPastor. (2021). Enhancing precision medicine: A big data-driven approach for the management of genomic data. *Big Data Research, Elsevier*.
17. M. Aliyari. (2021). Securing industrial infrastructure against cyber-attacks using machine learning and artificial intelligence at the age of industry 4.0. *Turkish Journal of Computer and Mathematics Education*.
18. Nazar Waheed, Xiangjian He, Muhammad Ikram, Muhammad Usman, Saad Sajid Hashmi, & Muhammad Usman. (2020). Security and privacy in IoT using machine learning and blockchain: Threats and countermeasures. *ACM Computing Survey, 53*(3).
19. Jagreet Kaur, & Dr. Kulwinder Singh Mann. (2021). Persuasive factors and weakness for security vulnerabilities in big IOT data in healthcare solution. *Journal of Physics, IOP Publishing*.
20. Jithin Jagannath, Nicholas Polosky, Anu Jagannath, Francesco Restuccia, & Tommaso Melodia. (2019). Machine learning for wireless communications in the Internet of Things: A comprehensive survey. *Ad Hoc Networks, Elsevier*.
21. Shakila Zaman, Khaled Alhazmi, Mohammed A. Aseeri, Muhammad Raisuddin Ahmed, Risala Tasin Khan, M. Shamim Kaiser, & Mufti Mahmud. (2021). Security threats and artificial intelligence based countermeasures for internet of things networks: A comprehensive survey. *IEEE Access*.
22. Mohamed Amine Ferrag, Othmane Friha, Leandros Maglaras, Helge Janicke, & Lei Shu. (2021). Federated deep learning for cyber security in the internet of things: Concepts, applications, and experimental analysis. *IEEE Access*.
23. Kaiyue Zhang, Xuan Song, Chenhan Zhang, & Shui Yu. (2022). Challenges and future directions of secure federated learning: a survey. *Frontiers of Computer Science, Springer*.