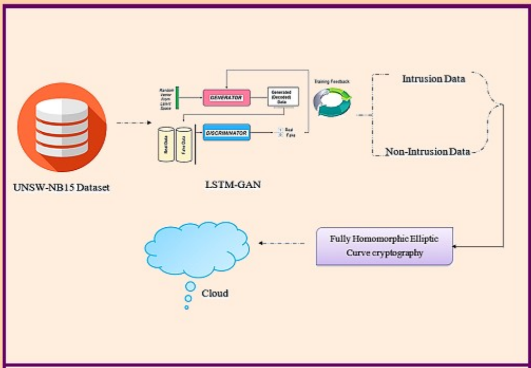
Volume 02, Number 03, 2024

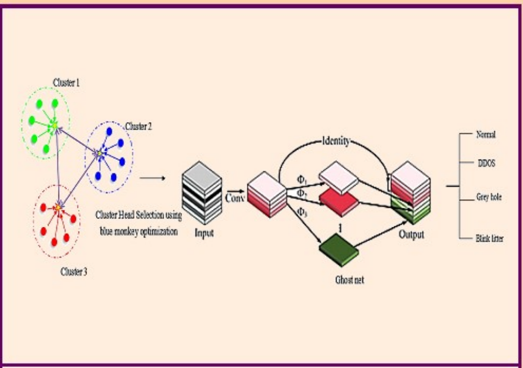
KITS PRESS

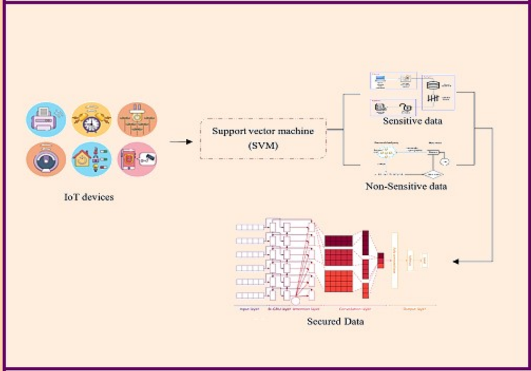
IJDSAI

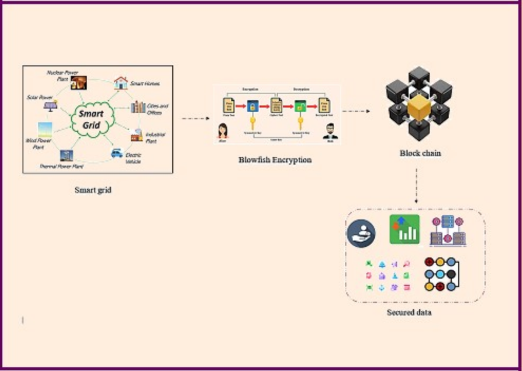
ISSN: 2584-1041

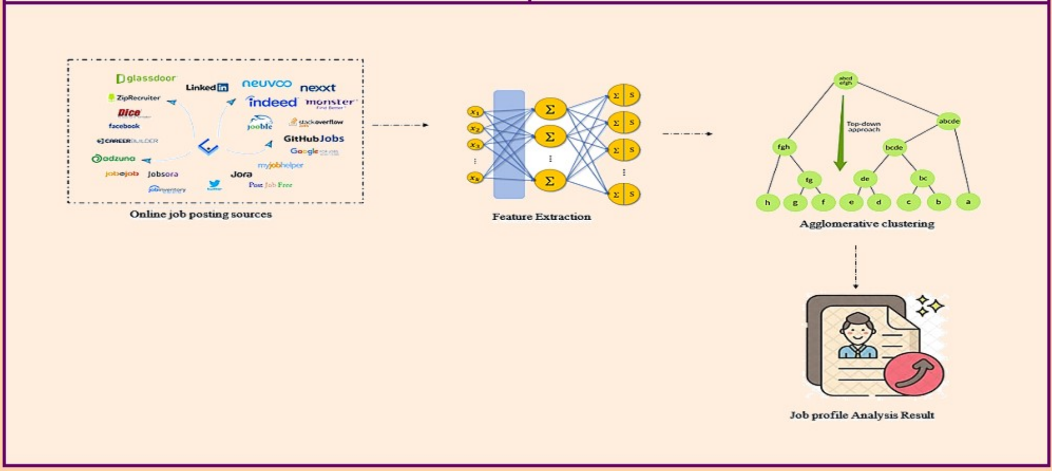
International Journal of Data Science and Artificial Intelligence









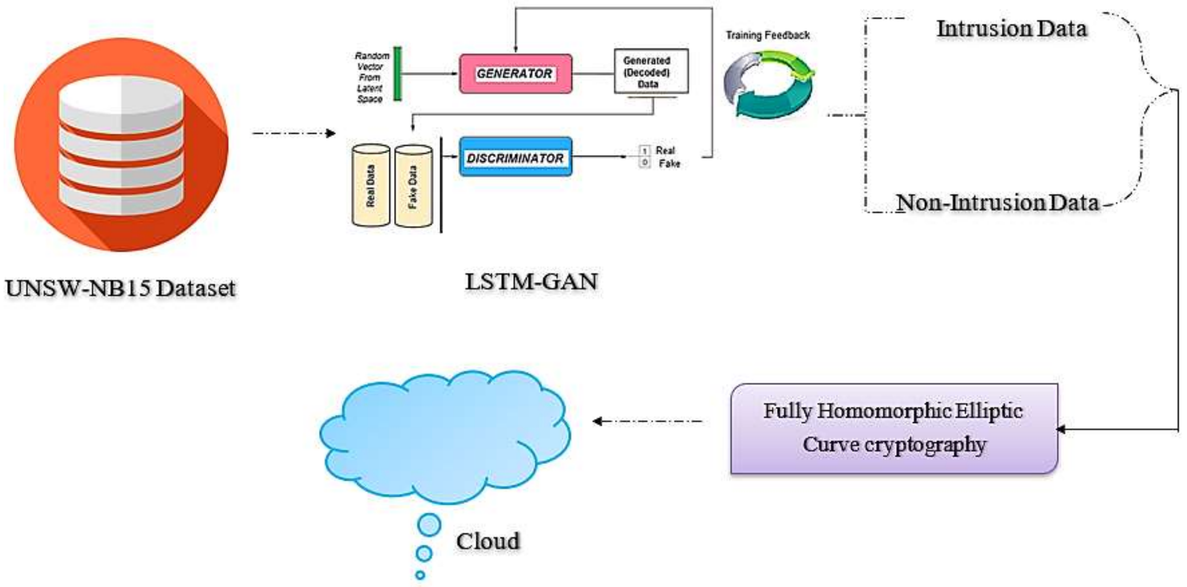


International Journal of Data Science and Artificial Intelligence

IJDSAI

1. DEEP LEARNING BASED LSTM-GAN APPROACH FOR INTRUSION DETECTION IN CLOUD ENVIRONMENT

Rajendran Arulappan Mabel Rose, Aarthi Gopalakrishnan and J. Vasuki

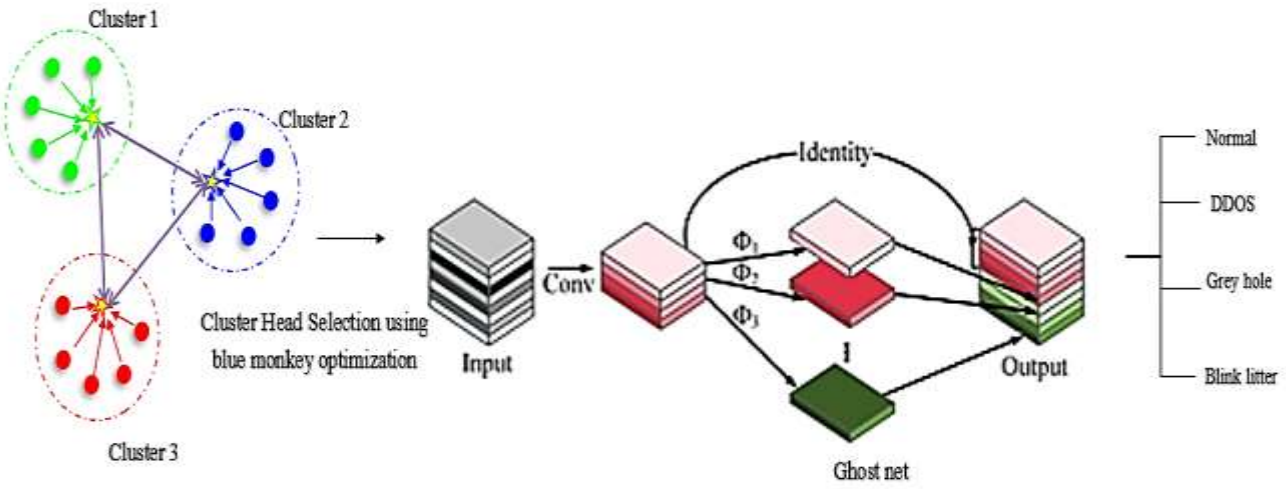


Abstract – Cloud computing is a rapidly growing technology paradigm with enormous potential. While cloud computing has many advantages, it also poses new security risks. Cloud computing security vulnerabilities have been identified as the most significant impediment to reaping its many benefits. When sensitive data and business applications are outsourced to a third party, these security concerns become critical. In an existing study, research has found that cloud-based intrusion detection systems (IDS) difficult, more time-consuming, and less secure are major issues. The paper proposes Intrusion Detection Systems and fully homomorphic elliptic curve cryptography (IDS-FHECC). LSTM-GAN is a deep learning algorithm that detects intrusion and non-intrusion data. To reduce computing complexity and improve security, FH-ECC is utilized to encrypt the input data. The encrypted data is processed using homomorphic operations such as implicit additions and multiplications. In the process of evaluating the proposed mechanism, different metrics like as accuracy, encryption, and decryption time are utilized. The proposed improves the overall accuracy by 8.5%, 9.8%, and 12.5% better than LOA, AODV, RLWE, and WOA respectively. The encryption time of the proposed method is 21.8%, 35.7%, and 39.2% decreased compared to existing LOA, AODV, RLWE, and WOA methods.

***Keywords* –** *Cloud computing, Intrusion detection systems, fully homomorphic encryption (FHE), Deep learning.*

2. SELECTIVE FORWARDING ATTACKS DETECTION IN WIRELESS SENSOR NETWORKS USING BLUE MONKEY OPTIMIZED GHOSTNETWORK

Jeyaraman Sathiamoorthy, M. Usha and P. Senthilraja



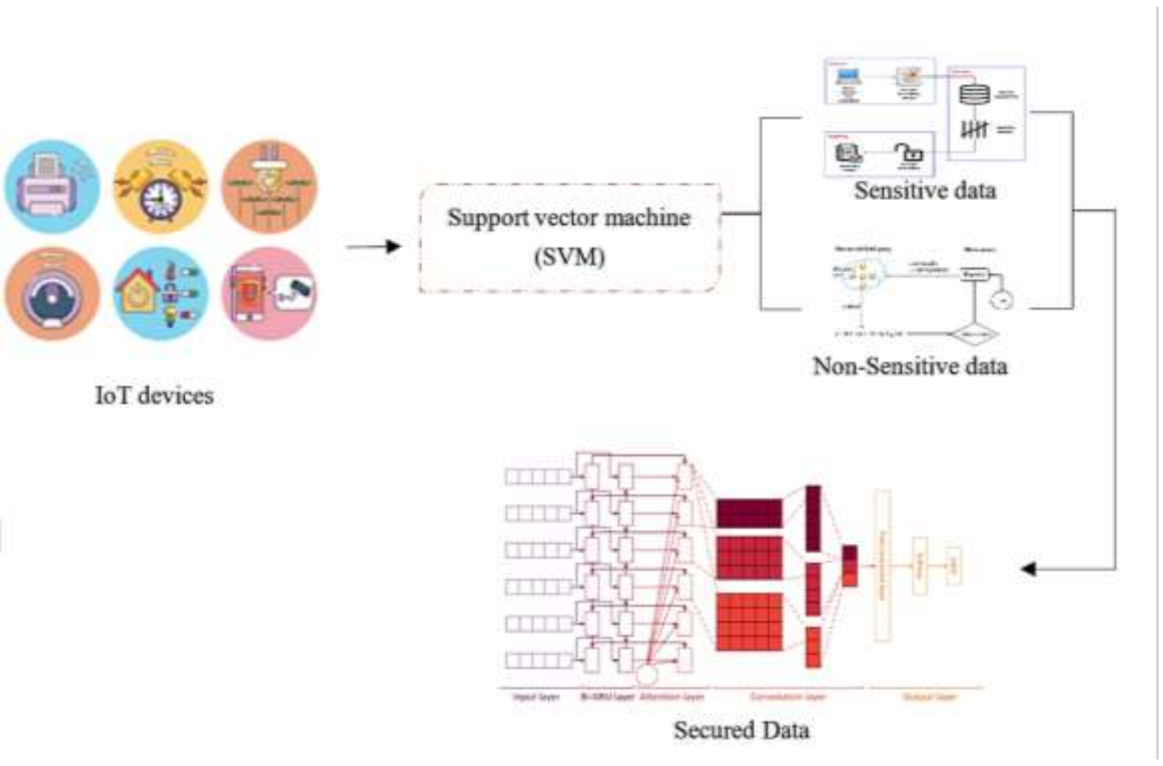
Abstract – Wireless Sensor Networks (WSNs) are increasingly the technology of choice due to their wide applicability in both military and civilian domains. The selective forwarding attack, one of the main attacks in WSNs, is the hardest denial-of-service attack to detect. The hostile nodes that initiate the selective forwarding attack will discard some or all of the data packets they receive. Numerous detection techniques for optional forwarding have been developed attacks are inaccurate or contain sophisticated algorithms, which is especially true when the attacker also uses other attacks like distributed denial of service, wormholes, and black holes to move through the network. To address these disadvantages, this research proposes a novel selective forwarding attack detection method based blue monkey-optimized ghost net (SAD-Ghost) method. To identify network threats, Blue Monkey optimization based on the hazard model is built in this case. A proposed technique to improve detection accuracy and minimize computation. The primary goal of the research is to develop a Selective Forwarding Attack Detection utilizing a blue Monkey optimized Ghost net to improve network lifetime. Initially, Blue Monkey optimization is used for optimal cluster head selection based on node degree and density. Moreover, cluster data are pre-processed using Tokenization, Normalization, and Reduction. The proposed method is utilized to detect intrusion in WSN and classify Normal, DDOS, Grey hole, and Blink litter. The experimental analysis demonstrates that the proposed method achieves a packet delivery rate of 97.6%, 95.3% and 90.50% and reduces energy consumption by 19.6%, 12.5% and 17.4% compared to existing clustering-based routing methods. Consequently, the proposed technique surpasses current methods in terms of network lifetime, energy efficiency, and packet delivery performance.

Keywords ***–*** *Wireless Sensor Networks, Intrusion Detection, Blue monkey, Selective forwarding attack detection.*

3. IOT-CENTRIC DATA PROTECTION USING DEEP LEARNING TECHNIQUE FOR PRESERVING SECURITY AND PRIVACY IN CLOUD

C. Senthil Singh, Sameena Naaz and G. Saranya

Abstract – A system of interconnected, The Internet of Things (IoT) is a term that refers to physical objects that may be connected online. Concerns over user privacy on the Internet of Things are growing as a result of the large amounts of personal data being gathered and exchanged there. IoT devices may increase productivity, accuracy, and financial gain in addition to reducing human intrusion, giving Internet of Things applications the most flexibility and convenience. Overhead of communications, security, and privacy, IoT is experiencing issues as well. As a result, protecting data has grown to be a difficult undertaking that must be handled carefully. This study offers a secure data security solution for preserving privacy in the cloud environment to address this crucial and difficult topic. It does this by efficiently separating the data by separating sensitive data from non-sensitive data with an SVM classifier and then using the data to partially decrypt and analyze, which increases the effectiveness of the model while ensuring security. The sensitive data was protected using Okamoto Uchiyama encryption. The model safely stores, analyzes, and shares data to ensure the system's safety and privacy. The novel method was compared to existing methodologies in terms of particular parameters like precision, accuracy, F1 score, and recall. When compared to other schemes, the proposed solution offers both high security and efficiency.

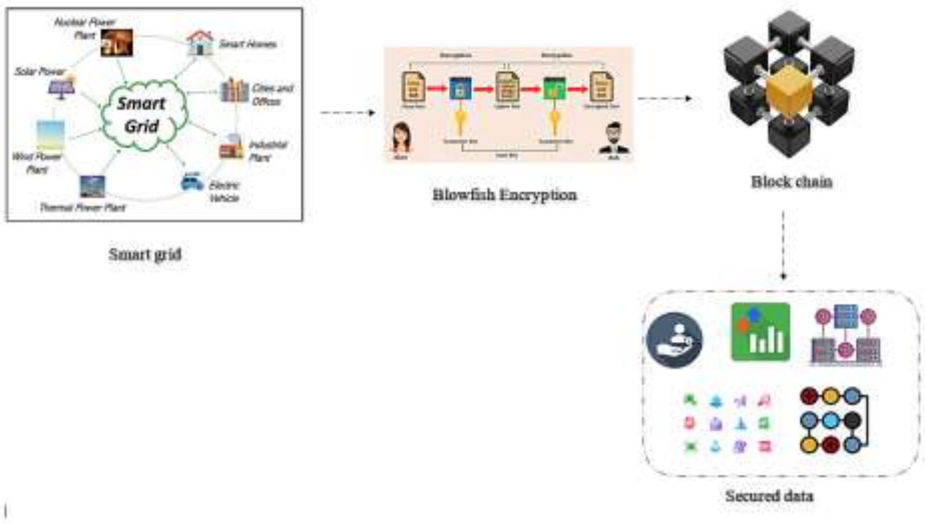


Keywords ***–*** *cloud computing, Internet of things, Support vector machine, Okamoto Uchiyama.*

4. BLOCK CHAIN ENABLED DATA SECURITY USING BLOWFISH ALGORITHM IN SMART GRID NETWORK

R. C. Ilambirai, S. Lourdu Jame and P.U. Poornima

Abstract – Smart Grid provides a reliable and efficient end-to-end delivery system. Data on each user's unique electricity consumption is given in real time. It also enables utilities to control and monitor the electrical system in real-time, helping them to reduce power outages. Users' privacy is a significant issue in addition to the usual security issues. Data about power usage may be used to infer private information about users by entities with access to the data. To solve these problems, a Block-Chain-based Secure Smart Grid Network (BCS-SGN) has been created, utilizing group signatures and covert channel permission to guarantee user validity. Initially, the data from the smart grid network will be



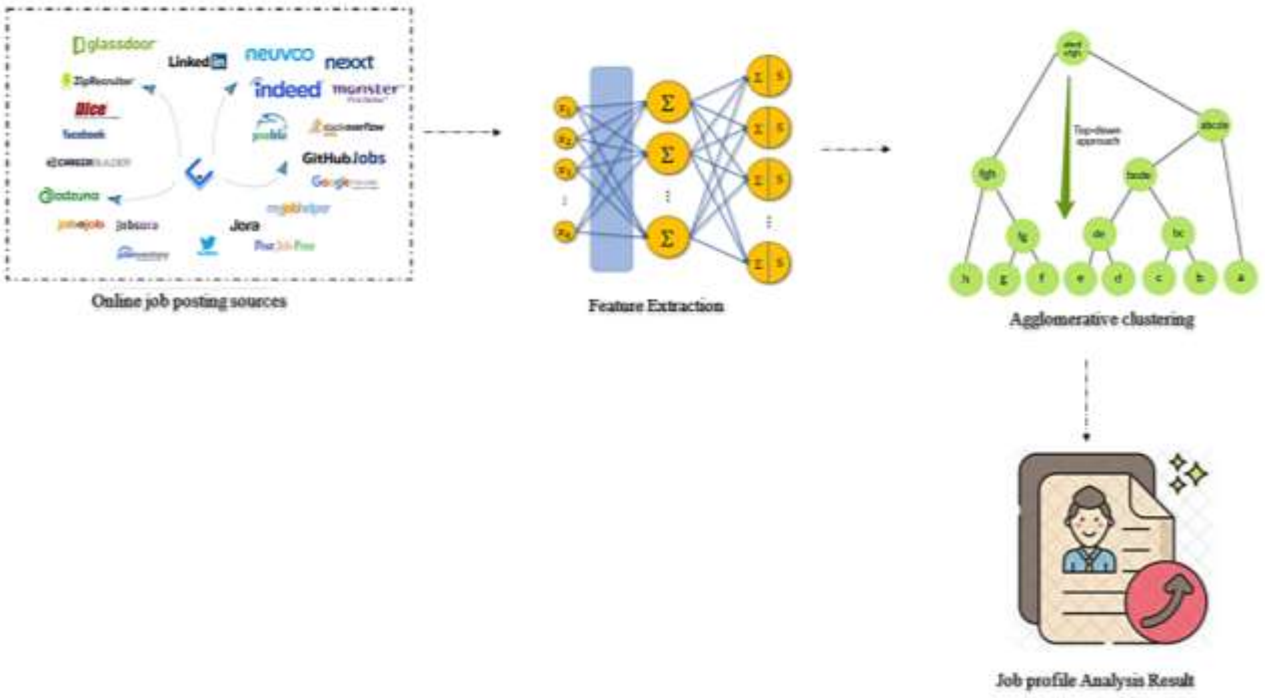
collected and encrypted using blowfish techniques. After encryption, the encrypted data will be stored in the block chain network, which also stores the transmission logs. In order to utilize the security controller, smart contracts are analyzed using smart grid devices. DES, AES, and BCS-SGN are widely used. Symmetric key algorithms are properly compared in this study. Additionally, the metrics of encryption/decryption time, computational time, and throughput are also compared. The percentage of the proposed method, BCS-SGN, is 20%, AES is 15%, and DES is 12%. These outcomes demonstrate that the proposed BCS-SGN outperforms other techniques.

Keywords – *Smart grid, Smart contract, Blowfish Algorithm, Block chain.*

5. SEMANTIC FEATURE ENABLED AGGLOMERATIVE CLUSTERING FOR INFORMATION TECHNOLOGY JOB PROFILE ANALYSIS

B. Jaison, R. Gladis kiruba and G. Sreetha

Abstract – The maintenance and implementation of computer systems are the core activities of information technology. Database administration and network architecture are also included in information technology. Professionals have access to a working environment that facilitates the setup of internal networks and the development of computer systems. There is an immediate need for a suitable approach to close the gap between supply and demand for IT workers. Extensive research into IT job profiles is crucial to meeting industry demands. Educational programs must identify the abilities that the industry requires to modernize its manufacturing. Semantic Feature-Enabled Agglomerative Clustering for Information Technology Job Profiling (SEA-IT) has been proposed to overcome these challenges. Semantic analysis is performed using a tree-like strategy. The most frequently used phrases and words from each cluster of IT professions were collected to demonstrate specific knowledge. Initially, the data from the online job posting sources will be collected and pre-processed using techniques such as stemming, normalization, text correction, removing stop words, and tokenization. Secondly, the pre-processed data can extract features using a bag of words. After feature extraction, the cluster is generated using an agglomerative algorithm to form an IT job analysis result, so that the knowledge and capabilities of IT professionals can be upgraded. The simulation findings, based on evaluation criteria and other statistical tests, demonstrated the suggested algorithm. Experiments demonstrated that SEA-IT functions well with a variety of descriptive methodologies and is independent of the dataset's dimensions.



Keywords – *Information Technology, Preprocessing, bag of words, agglomerative algorithm.*

International Journal of Data Science and Artificial Intelligence (IJDSAI)
Volume 02, Issue 03, May – June (2024)

RESEARCH ARTICLE

DEEP LEARNING BASED LSTM-GAN APPROACH FOR INTRUSION DETECTION IN CLOUD ENVIRONMENT

Rajendran Arulappan Mabel Rose [1, *], Aarthi Gopalakrishnan[2] and J. Vasuki [3]

[1] Department of Computer Science and Business System, Panimalar Engineering College, Thiruverkadu, Tamil Nadu 600077, India.

[2] Department of Computer Science and Engineering, B.S. Abdur Rahman Crescent Institute of Science and Technology, Chennai, India.

[3] Department of Computer Science and Engineering, Research Point India private limited, Nagercoil, India.

Corresponding e-mail: mabel.nidhu@gmail.com

Abstract – Cloud computing is a rapidly growing technology paradigm with enormous potential. While cloud computing has many advantages, it also poses new security risks. Cloud computing security vulnerabilities have been identified as the most significant impediment to reaping its many benefits. When sensitive data and business applications are outsourced to a third party, these security concerns become critical. In an existing study, research has found that cloud-based intrusion detection systems (IDS) difficult, more time-consuming, and less secure are major issues. The paper proposes Intrusion Detection Systems and fully homomorphic elliptic curve cryptography (IDS-FHECC). LSTM-GAN is a deep learning algorithm that detects intrusion and non-intrusion data. To reduce computing complexity and improve security, FH-ECC is utilized to encrypt the input data. The encrypted data is processed using homomorphic operations such as implicit additions and multiplications. In the process of evaluating the proposed mechanism, different metrics like as accuracy, encryption, and decryption time are utilized. The proposed improves the overall accuracy by 8.5%, 9.8%, and 12.5% better than LOA, AODV, RLWE, and WOA respectively. The encryption time of the proposed method is 21.8%, 35.7%, and 39.2% decreased compared to existing LOA, AODV, RLWE, and WOA methods.

Keywords *– Cloud computing, Intrusion detection systems, fully homomorphic encryption (FHE), Deep learning.*

1. INTRODUCTION

Cloud computing has quickly become a globally known virtual machine in recent years, with on-demand delivery of apps, platforms, and computing infrastructures likely to play a vital role in the upcoming Internet of Services [1]. Users can keep data and run programmes on top of a virtual computing infrastructure provided by cloud computing [2]. Computing in the cloud allows easy, on-demand access to computer resources [3]. Cloud-based resources can be accessed by organizations on a need-to-know basis. Cloud computing facilitates the analysis of large volumes of data by making use of shared computer resources while being able to easily adapt to changes in data volume and variety.

Cloud service providers (CSPs) provide users with on-demand access to these services while they invest in the infrastructure and computational backbone [4]. Data storage and computing are the two categories of cloud computing services. Cloud service users [5] don't need to know where their data is stored in the cloud or which machines in their network are performing their computations. It is crucial to ensure that cloud data is secure while communicating and uncheckable when building up the cloud architecture. To do this, a wide variety of authentication and encryption methods must be incorporated, which are constantly evolving to deal with new risks and problems.

While encryption protects data at rest, the data can be lost if the encryption key is lost. It is essential to develop cryptographic solutions that can conduct calculations on encrypted data without decryption to prevent malicious attacks on the cloud. Cryptographic algorithms keep communication between individuals, groups, and organizations as private as possible [6]. Cryptographic algorithms play a critical role in data or information security, lowering or eliminating the risk of data leaks [7, 8].

A growing number of factors are contributing to data transmission security in cloud computing. Increasingly, people are concerned about the security of the data they store in the cloud because of frequent hacking incidents. There are several security limitations and difficulties that arise during data transmission in a cloud environment. It is therefore essential to establish a strong intrusion detection system (IDS) to detect and avoid attacks at an early stage. Therefore, intrusion detection systems (IDS) are now an essential part of computer and network security.

The key contribution of this paper is to secure communication in the cloud using Intrusion Detection Systems and fully homomorphic Elliptic Curve Cryptography (IDS-FHECC).

- In this section, Intrusion Detection Systems and fully homomorphic elliptic curve cryptography (IDS-FHECC) have been proposed.
- The proposed system has three phases namely pre-processing, intrusion detection classification, and data encryption.
- Initially, the UNSW-NB15 dataset is pre-processed using tokenization, Repeated Word Removal, and dimension reduction. Intrusion detection is classified using a deep learning framework namely LSTM-GAN. The classified non-intrusion data is encrypted using a Fully Homomorphic-Elliptic Curve Cryptography.
- In the process of evaluating the proposed mechanism, different metrics like as accuracy, encryption, and decryption time are utilized. The proposed improves the overall accuracy by 8.5%, 9.8%, and 12.5% better than LOA, AODV, RLWE, and WOA respectively.

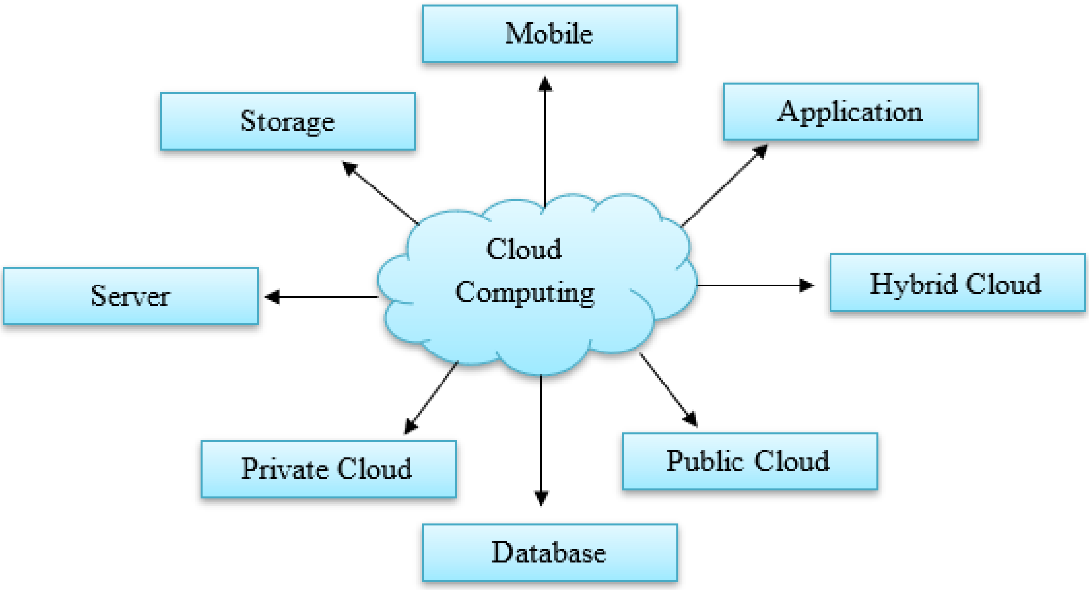


Figure 1. Cloud Computing Models

The research's remaining portion is illustrated in the manner that follows. Section II explains the literature review. Section III included an explanation, the corresponding algorithm, and examples of the suggested techniques. Section IV contains the performance results and analyses. In Section V, conclusions and additional work were finished.

2. LITERATURE SURVEY

In 2023, Kasongo, S.M., [9] A new Intrusion Detection System (IDS) utilizing the Tree-CNN hierarchical algorithm and featuring the Soft-Root-Sign (SRS) activation function has been introduced. The model reduces the training time of the generated model for detecting DDoS, Infiltration, Brute Force, and Web attacks. Additionally, they test whether the proposed model performs well in binary and multiclass classifications by examining the effects of the number of nerve cell and erudition rates. RNN-IDS performs better in binary and multiclass classification than classical machine learning approaches in the experiments, showing that it can model a classification model with high accuracy.

In 2022, Alghamdi, S.A., [10] proposed a novel framework named Trust-Aware Intrusion Detection and Prevention System (TA-IDPS) aimed at safeguarding networks from potential adversaries. Combining symmetric and asymmetric techniques will provide great security, and every user will have an encrypted private key that may be used by multiple people simultaneously. In addition to hybrid cryptography's security the key's security will be increased by this method.

In 2020, Adil M. et al [11] A novel lightweight anonymous authentication technique using MAC-based AODV was introduced to address the identified problem, specifically mitigating the threat of black-hole attacks. It is capable of identifying hostile nodes. In addition, it generates all probable pathways with their trust lengths, and the most trusted path is chosen for routing using AODV as the safest and optimal path. Simulated results show that this tactic is effective for identifying black hole attackers, calculating trust length ratios, and analyzing performance metrics. Consequently, the proposed method is more important than relying on neighbour nodes for trust.

In 2023, Salama. R, et al. [12] presented the SecureLR framework for protecting sensitive patient data by using logistic regression and hybrid cryptographic techniques incorporating secure hardware and homomorphic encryption. By using this technique, logistic regression can be used in the real world to analyze data related to patients, such as health data.

In 2023, Riya K.S, et al [13] proposed safe access in smart city applications, and mobile cloud computing was used. Using an efficient anonymous mutual authentication mechanism the suggested approach enables a mobile user to utilize a unique private key to access a variety of services via distinct service providers. Compared to conventional systems, the suggested technique is more efficient in terms of computational overhead.

In 2023, Patil.S et al. [14] proposed a homomorphic encryption protocol based on RLWE for message management and user authentication. In the proposed encryption protocol, identity values, user information, cloud identification values, and gateway identification values are specified. The investigation utilized a comparison analysis of computation space complexity and time to ensure that the suggested communication protocol delivers robust security and comparable efficiency. Confidentiality and decoding were used in the research.

In 2022, Balashunmugaraja, B, et al [15] proposed Whale Optimization (JWO) is a hybrid of the Whale optimization algorithm (WOA) and algorithm that uses homomorphic encryption to start safe cloud data transmission. A suggested JWO beats the existing techniques with scores of 0.720, 0.822, and 0.722 in well-being, BD, and accuracy, correspondingly.

3. PROPOSED METHODOLOGY

In this section, the Intrusion Detection System and fully homomorphic elliptic curve cryptography (IDS-FHECC) have been proposed. The proposed system has three phases namely pre-processing, intrusion detection classification, and data encryption. Initially, the UNSW-NB15 dataset is pre-processed using data normalization, label encoding, and removing normal traffic. Intrusion detection is classified using a deep learning framework namely LSTM-GAN. Using fully homomorphic-elliptic curve cryptography, the classified non-intrusion data is encrypted. Figure 2 illustrates the general suggested process.

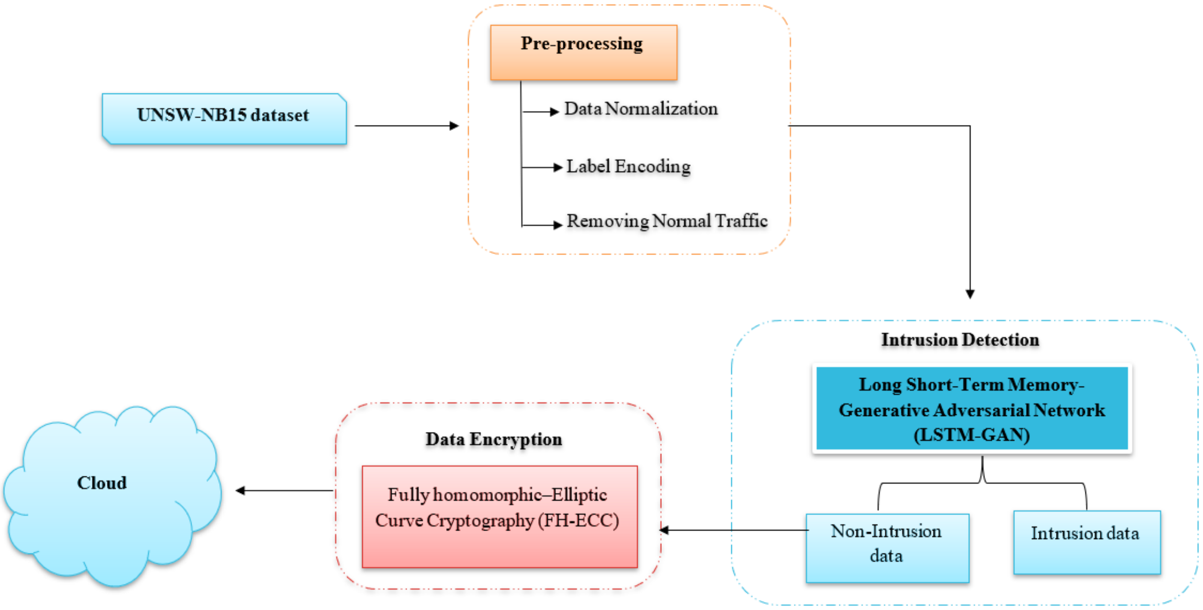


Figure 2. Overall IDS-FHECC Block diagram

3.1. Pre-processing

In this paper, the UNSW-NB15 dataset is used as input data. Different kinds of procedures, including data normalization, label encoding, and removing normal traffic, are performed on the incoming data during the pre-processing step. To achieve data normalization, each attribute's values must be normalized and should have 0 as a minimum number and a maximum number of 1. Attack labels are provided as string values for the multi-class labels in the dataset. Information packets reflecting typical network usage from both datasets are ignored for multiclass classification since they make up the majority of the data. Only attack data streams**.**

3.2. Intrusion Detection using LSTM-GAN

Recurrent neural networks (RNNs) of the artificial variety are similar to the LSTM. This network receives different inputs than conventional recurrent neural networks. It can analyze entire streams of data at a single point. Since there may be unidentified lags between significant events in data from time series, LSTM networks are a good fit for categorization, analysis, and forecasting. Problems with the breakdown and disappearance of gradients have been addressed using LSTMs. This indicates that it ought to take into account every value in the signal and that a memory-based model ought to be employed. Thus, new ECG signals can be created using Long Short-Term Memory GAN (LSTM-GAN). Because of its structure, LSTM attempts to correlate the upcoming values with the past values, remembering previous knowledge. This allows for the processing of data such as frequency variations, amplitude values, and signal rise and fall. Figure 3 below depicts the LSTM-GAN model's architecture, which was used in this investigation.

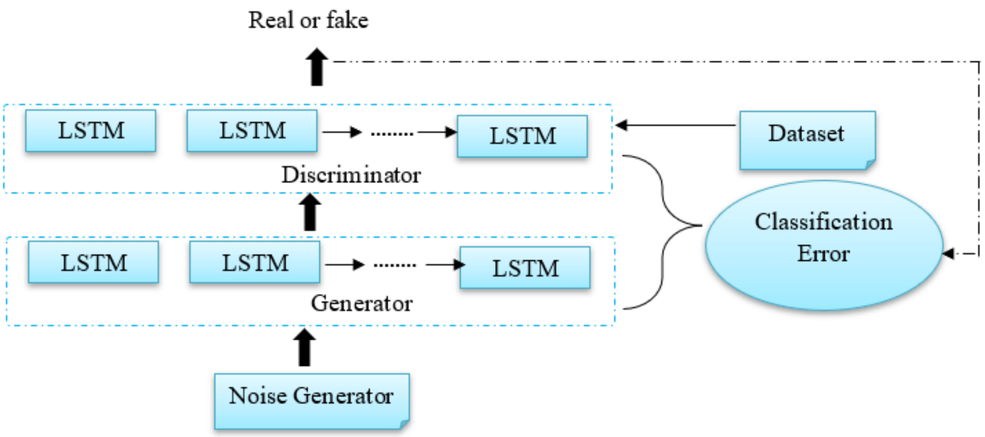


Figure 3. LSTM-GAN Architecture

There are two distinct loss parameters defined for the generator and discriminator, and these functions are employed in tandem.

$$l_d = E\big(d_{(X)}, 1\big) + E\big(d\big(g(z)\big), 0\big) \quad (1)$$

$$l_g = E((d\big(g(z)\big), 1) \quad (2)$$

$$\text{H (P, Q)} = E_{X \sim p(x)}[-logQ(X)] \quad (3)$$

The presence of LSTM modules within the Generator and Discriminator units distinguishes the LSTM-GAN model from the traditional GAN. LSTM-GAN aims to reduce generator loss, increase discriminator loss, and achieve Nash equilibrium, similar to classical GAN.

3.3. Data encryption using FH-ECC:

For secure data storage, a technique called fully homomorphic-elliptic curve cryptography (FH-ECC) is created. The FH is used as an asymmetric method in this study, and the functionalities of the FH are changed by the ECC.

Algorithm for Encryption in FH-ECC:

Input: Private Key (P_r)

Output: Encrypted data (E_d)

```
1.  Create Public key ($P_u$)
2.  Public key =$P_r$* $G_f$, where $G_f$ denotes generator function
3.  Create the cipher text
4.  Cipher text=$P_t$* $G_f$
5.  Create encrypted data
6.  $E_d$= ($P_t$*$P_u$) + ($d_j$, $P_c$), where $P_c$ denotes the point of the curve
7.  Consider $E_d$ is an encrypted data for FH scheme is applicable
8.  $E_d$ = ($P_t$+$P_u$) $\sum_{j=1}^{f}$.($d_j$, $P_c$)
9.  Implement Homomorphic scheme using multiplication property FHE→F $_{(*/)}$
10. For encrypted data, Calculate 3 coefficient a, b, c
11. $F_a$ = FHE (a1, a2, a3)
12. Calculate $F_b$ = FHE (b1, b2, b3)
13. Create $F_c$ = FHE ($d_{Fb}$, c)
14. Finally, the encrypted data $F_c$ ($E_d$) is uploaded to the cloud.
```

Encryption: It remains the process of turning plain text into the encrypted text to improve security. An often-used cloud security mechanism called the ECC offers protection according to the complexity of specific challenges. One advantage of this technique is that it just depends on the listed table and key. Moreover, it provides improved data solutions and enables safe key exchange across communication devices. Multiplication and addition on encrypted data, as well as arbitrary data calculations, are made possible via FH operations. To safely share data in the cloud, this study uses the FH-ECC technique.

Decryption: Data is decrypted when it is received by the user (e.g., employees or students). The cipher text is extracted from the encrypted data using FH processes, and the cipher text is then converted back to the primary text by decrypting it with the ECC algorithm.

4. RESULT AND DISCUSSION

The suggested IDS-FHECC performance is estimated using several known approaches in this part. The proposed job is carried out with the help of Java-running software and cloud computing. The text documents are obtained experimentally from the NSL-KDD data set, which includes intrusion and non-intrusion data.

4.1. Performance Analysis

The performance evaluation in this study is determined utilizing performance metrics including sensitivity (Sen), accuracy (A), and specificity (Spec), as stated in the equations below.

$$Sensitivity = \frac{T_P}{T_P + F_N} \quad (1)$$

$$Specificity = \frac{T_N}{T_N + F_P} \quad (2)$$

$$Accuracy = \frac{T_P + T_N}{Total\ no.of\ samples} \quad (3)$$

For Equations (1), (2), and (3), TP, TN, FN, and FP, each stand for genuine positives, phony positives, and phony negatives. In a TP result, intrusion documents have been correctly classified, whereas in a TN result, nonintrusive documents have been classified appropriately.

4.2. Comparative Analysis

The efficiency of the suggested technique is demonstrated in the graph below, which compares training and testing accuracy over numerous epochs for intrusion detection. Figure 4 shows how, after 45 epochs, the training accuracy gradually approaches 100%. In 50 epoch, numbers, convergence is faster and more efficient, as shown.

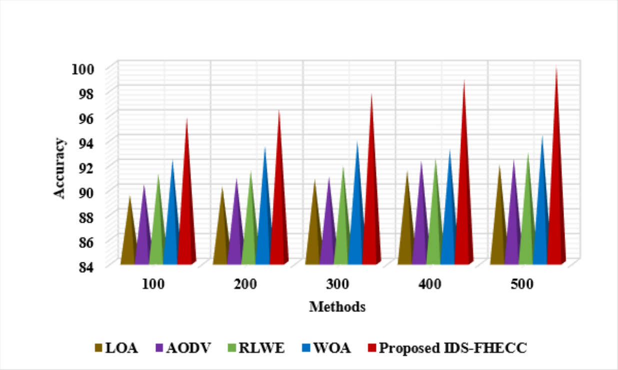


Figure 4. Comparative Analysis of proposed Vs Existing

In terms of accuracy, the presented approach outperforms all other algorithms shown in figure 4. As a consequence, the suggested model achieves accuracy of 99.87 percent.

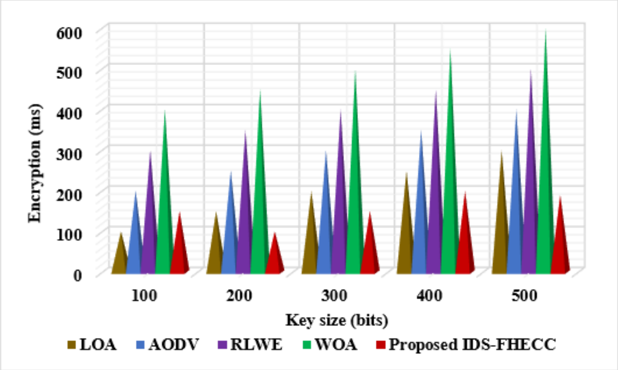


Figure 5. Comparative Analysis of Proposed Vs Existing Methods

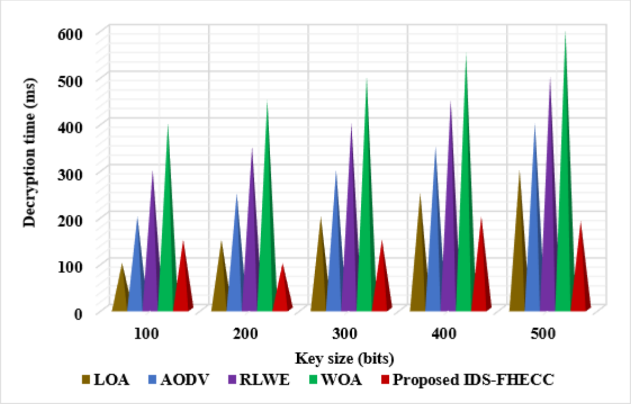


Figure 6. Comparative Analysis of Proposed Vs Existing Methods

Figure 5 shows the encryption times of the existing LOA, AODV, RLWE, WOA, and suggested IDL-FHECC algorithms. An increase in key size (bits) can enhance encryption time, according to this study. The analysis indicates that the IDL-FHECC methodology requires shorter encrypting duration than alternative methods. As a result, homomorphic operations are performed on separated formatted data that was encrypted.

According to Figure 6, the decryption times of existing LOA, AODV, RLWE, WOA, and suggested IDL-FHECC algorithms vary with key size (bits). By increasing the key size linearly, the decryption time can be enhanced.

5. CONCLUSION

The paper proposes Intrusion Detection Systems and fully homomorphic elliptic curve cryptography (IDS-FHECC). LSTM-GAN is a deep learning algorithm that detects intrusion and non-intrusion data. In this section, Intrusion Detection Systems and fully homomorphic elliptic curve cryptography (IDS-FHECC) have been proposed. The proposed system has three phases namely pre-processing, intrusion detection classification, and data encryption. Initially, the UNSW-NB15 dataset is pre-processed using tokenization, Repeated Word Removal, and dimension reduction. Intrusion detection is classified using a deep learning framework namely LSTM-GAN. The classified non-intrusion data is encrypted using a Fully Homomorphic-Elliptic Curve Cryptography. In the process of evaluating the proposed mechanism, different metrics like as accuracy, encryption, and decryption time are utilized. The proposed improves the overall accuracy by 8.5%, 9.8%, and 12.5% better than LOA, AODV, RLWE, and WOA respectively. The encryption time of the proposed method is 21.8%, 35.7%, and 39.2% decreased compared to existing LOA, AODV, RLWE, and WOA methods. The proposed methodology achieves the best results and is extremely safe, decreasing the complexity of encryption and decryption.

CONFLICTS OF INTEREST

This paper has no conflict of interest for publishing.

FUNDING STATEMENT

Not applicable.

ACKNOWLEDGEMENTS

The author would like to express his heartfelt gratitude to the supervisor for his guidance and unwavering support during this research for his guidance and support.

REFERENCES

[1] S.A. Bello, L.O. Oyedele, O.O. Akinade, M. Bilal, J.M.D. Delgado, L.A. Akanbi, A.O. Ajayi, and H.A. Owolabi, "Cloud computing in construction industry: Use cases, benefits and challenges", *Automation in Construction*, vol. 122, pp. 103441, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[2] B. Alouffi, M. Hasnain, A. Alharbi, W. Alosaimi, H. Alyami, and M. Ayaz, "A systematic literature review on cloud computing security: threats and mitigation strategies", *IEEE Access*, vol. 9, pp. 57792-57807, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[3] T. Alam, "Cloud Computing and its role in the Information Technology", *IAIC Transactions on Sustainable Digital Innovation (ITSDI)*, vol. 1, no. 2, pp. 108-115, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[4] R. Maeser, “Analyzing CSP Trustworthiness and Predicting Cloud Service Performance”, *IEEE Open Journal of the Computer Society*, vol. 1, pp. 73-85, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[5] F. Nawaz, M.R. Asadabadi, N.K. Janjua, O.K. Hussain, E. Chang, and M. Saberi, “An MCDM method for cloud service selection using a Markov chain and the best-worst method”, *Knowledge-Based Systems*, vol. 159, pp. 120-131, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[6] R. Abid, C. Iwendi, A.R. Javed, M. Rizwan, Z. Jalil, J.H. Anajemba, and C. Biamba, “An optimised homomorphic CRT-RSA algorithm for secure and efficient communication”, *Personal and Ubiquitous Computing*, pp. 1-14, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[7] S. Tayal, N. Gupta, P. Gupta, D. Goyal, and M. Goyal, “A review paper on network security and cryptography”, *Advances in Computational Sciences and Technology*, vol. 10, no. 5, pp. 763-770, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[8] I. Bhardwaj, A. Kumar, and M. Bansal, “A review on lightweight cryptography algorithms for data security and authentication in IoTs”, In *2017 4th International Conference on Signal Processing, Computing and Control (ISPCC)*, pp. 504-509, 2017. IEEE. [CrossRef] [Google Scholar] [Publisher Link]

[9] S.M. Kasongo, “A deep learning technique for intrusion detection system using a Recurrent Neural Networks based framework”, Computer Communications, vol.199, pp.113-125, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[10] S.A. Alghamdi, “Novel trust-aware intrusion detection and prevention system for 5G MANET–Cloud”, International Journal of Information Security, vol. 21, no. 3, pp. 469-488, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[11] M. Adil, R. Khan, M.A. Almaiah, M. Al- Zahrani, M. Zakarya, M.S. Amjad, and R. Ahmed, “MAC-AODV based mutual authentication scheme for constraint-oriented networks”, Ieee Access, vol. 8, pp. 44459-44469, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[12] R. Salama, S. Al-Turjman, C. Altrjman, F. Al-Turjman, R.O. Prakash, S.P. Yadav, and S. Vats, “Authentication using Biometric Data from Mobile Cloud Computing in Smart Cities”, In 2023 3rd International Conference on Advancement in Electronics & Communication Engineering (AECE) 445-448 (2023). IEEE. [CrossRef] [Google Scholar] [Publisher Link]

[13] K.S. Riya, R. Surendran, C.A. Tavera Romero, and M.S. Sendil, “Encryption with User Authentication Model for Internet of Medical Things Environment”, Intelligent Automation & Soft Computing, vol. 35, no. 1, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[14] S. Patil, and R. Patil, Jaya-based, “CAViaR: Hadamard product and key matrix for privacy preservation and data sharing in cloud computing environment”, International Journal of Grid and Utility Computing, vol. 14, no. 4, pp. 389-399, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[15] B. Balashunmugaraja, and T.R. Ganeshbabu, “{Privacy preservation of cloud data in business application enabled by multi-objective red deer-bird swarm algorithm”, Knowledge-Based Systems, vol. 236, pp.107748, 2022. [CrossRef] [Google Scholar] [Publisher Link]

AUTHORS

Rajendran Arulappan Mabel Rose received her B.E. degree and M.E. degree in Computer Science and Engineering from Anna University, Chennai, India. She started her career as Lecturer and has 13 years of experience. Currently she is working as Assistant Professor in S.A. Engineering College, Chennai. Her research interests include Wireless Networks and Cloud Computing. She is a lifetime member of MISTE.

Aarthi Gopalakrishnan Completed her B.Tech in Information Technology and M.Tech in Computer and Communication from the Anna University in 2011. At present, she is pursuing Ph.D from B.S.Abdur Rahman Crescent Institute of Science and Technology. She has published/presented several research papers in National/ International Conferences.

J. Vasuki she was born in Kanyakumari District, Tamil Nadu, India in 1999. She received her BE degree in computer science and engineering from Arunachala college of engineering, Manavilai, Anna University, India in 2021. Currently she is a Research and Development Engineer in Research Point India private limited, Nagercoil, India. Her interested research area is image processing, machine learning and deep learning.

Arrived: 15.04.2024
Accepted: 02.06.2024

International Journal of Data Science and Artificial Intelligence (IJDSAI)
Volume 02, Issue 03, May – June (2024)

IJDSAI

RESEARCH ARTICLE

SELECTIVE FORWARDING ATTACKS DETECTION IN WIRELESS SENSOR NETWORKS USING BLUE MONKEY OPTIMIZED GHOST NETWORK

Jeyaraman Sathiamoorthy [1, *], M. Usha [2] and P. Senthilraja [3]

[1] Professor, Department of Computer science and Design, RMK Engineering College (Autonomous), RSM Nagar, Kavarapettai, Gummidipoondi Taulk, Thiruvallur District-601206, India.
[2] Professor and Assistant director, Department of Master of Computer Application, MEASI Institute of Information Technology, Royapettah, Chennai 600014, India.
[3] Assistant Professor, Department of Computer Science Engineering (Artificial Intelligence & Machine Learning), B.V. Raju Institute of Technology, Narsapur, Tuljaraopet, Telangana 502313, India.

Corresponding e-mail: jsathyam74@gmail.com

Abstract – Wireless Sensor Networks (WSNs) are increasingly the technology of choice due to their wide applicability in both military and civilian domains. The selective forwarding attack, one of the main attacks in WSNs, is the hardest denial-of-service attack to detect. The hostile nodes that initiate the selective forwarding attack will discard some or all of the data packets they receive. Numerous detection techniques for optional forwarding have been developed attacks are inaccurate or contain sophisticated algorithms, which is especially true when the attacker also uses other attacks like distributed denial of service, wormholes, and black holes to move through the network. To address these disadvantages, this research proposes a novel selective forwarding attack detection method based blue monkey-optimized ghost net (SAD-Ghost) method. To identify network threats, Blue Monkey optimization based on the hazard model is built in this case. A proposed technique to improve detection accuracy and minimize computation. The primary goal of the research is to develop a Selective Forwarding Attack Detection utilizing a blue Monkey optimized Ghost net to improve network lifetime. Initially, Blue Monkey optimization is used for optimal cluster head selection based on node degree and density. Moreover, cluster data are pre-processed using Tokenization, Normalization, and Reduction. The proposed method is utilized to detect intrusion in WSN and classify Normal, DDOS, Grey hole, and Blink litter. The experimental analysis demonstrates that the proposed method achieves a packet delivery rate of 97.6%, 95.3% and 90.50% and reduces energy consumption by 19.6%, 12.5% and 17.4% compared to existing clustering-based routing methods. Consequently, the proposed technique surpasses current methods in terms of network lifetime, energy efficiency, and packet delivery performance.

Keywords *– Wireless Sensor Networks, Intrusion Detection, Blue monkey, Selective forwarding attack detection.*

1. INTRODUCTION

Wireless sensor networks (WSN) are regarded as one of the most important research subjects [1]. WSN has multiple uses in industries such as telecommunications, the military, healthcare, research, and agriculture. They are using networks to identify natural calamities like earthquakes, flooding, and volcanoes [2]. Many security risks have been introduced throughout the implementation and deployment phases of WSNs due to their broad use [3]. WSN face unique challenges such as limited storage, computing power, and battery life, making them vulnerable to attacks [4].

Altogether, people depend on networking networks to deliver fresh suggestions, solutions to their problems, and assistance in fulfilling their fundamental needs [5]. The most recent and frequently utilized technology Sensors that enable users to get data remotely and use it for a specific purpose are examples of advancements [6]. The Internet of Things (IoT) is a growing field of technology that uses sensors [7]. Researchers are increasingly interested in WSNs, as evidenced by a rise in research papers [8]. While WSN has benefits, it can also be vulnerable to DoS attacks due to security flaws.

Several security risks that could lead to data breaches are possible for users to encounter when utilizing WSN apps. Researchers have been working on new security methods to prevent DoS assaults from being successful [9]. Numerous technical developments have aided in the creation of creative strategies for sneaking in and thwarting such attacks. However, deep learning has developed the best defenses against these kinds of security threats and denial-of-service attacks. DoS attacks overwhelm a service, preventing it from offering services to other users [10]. A DoS attack overwhelms your site or infrastructure with traffic from several sources, typically barring access for some time.

One of the services that protects websites from DoS attacks is Cloudflare. It could be difficult to defend against DoS assaults. There is nearly no way to stop the flow of that torrent of illegal content because it is coming at you from everywhere on the Internet and around the world. It faced many security challenges due to limited resources, insufficient infrastructure, and a high volume of WSN usage. DoS attacks frequently target the World Wide Web (WWW), and stopping them is no easy task. To overcome these challenges Selective forwarding Attack Detection utilizing blue monkey optimized Ghost net (SAD-Ghost) has been proposed. These are the primary benefits of the suggested methodology:

- The primary goal of the research is to develop a Selective Forwarding Attack Detection utilizing a blue Monkey optimized Ghost net to improve network lifetime.
- Initially, Blue Monkey optimization is used for optimal cluster head selection based on node degree and density. Moreover, cluster data are preprocessed using Tokenization, Normalization and Reduction.
- The proposed method utilizes Ghost net to detect intrusion in WSN and classify normal, DDos Grey hole, Blink litter.
- As a result, the proposed techniques outperform existing methods based on network lifetime, alive nodes, and residual energy.

The following examples illustrate the remaining section of this study. Section 2 explains the literature review. In Section 3, the proposed SAD-GHOST is displayed together with an explanation and the associated algorithm. Section 4 offers the performance results and their analysis. Section 5 includes conclusions and suggestions for future work.

2. LITERATURE SURVEY

In 2020, Premkumar and Sundararajan et al. [11] Offered a deep learning-based defense mechanism (DLDM) to detect and isolate DoS attacks during the data forwarding phase (DFP). DoS attacks like flooding, homing, jamming, and weariness may now be identified with greater accuracy owing to a revolutionary approach that has been documented in studies. The system in the simulation shows extremely high levels of detection, throughput, packet delivery ratio, and accuracy. It also minimizes the number of erroneous alerts and wasted energy.

In 2020, Asad et al. [12] offered a novel feed-forward back-propagation-based deep neural network detection method for accurately identifying several of the proposed neural network architecture that can precisely identify and exploit the key high-level characteristics of packet forays in application layer DDoS attacks on a cutting-edge dataset that includes many DDoS attacks. Therefore, to overcome these obstacles or risks, they are trying to use artificial intelligence technology to identify them.

In 2022 Salmi and Oughdir et.al [13] suggested using CNN-LSTM to identify and categorize DoS intrusion attempts as floods, TDMA, blackholes, normal, or gray holes. The research used "A state-of-the-art dataset for a computer-generated WSN-DS. The dataset includes a built-in model" that accurately classified the provided attacks with 99% accuracy, showing promising results in attack detection.

In 2022, Salmi and Salim et.al. [14] proposed a method for noticing DOS attacks in WSN using a CNN-LSTM model. In this type of attack, the endpoints flood a specific target with traffic, preventing genuine users from accessing its services. The CNN-LSTM model developed was evaluated over 25 epochs and achieved accuracy, precision, and recall scores of 0.943, 0.958, and 0.921, respectively.

In 2021, Wazir Ali and Ahmad et al. [15] proposed the use of machine learning to sense flooding, and gray holes, to ensure the security of WSN, they must prevent black hole-distributed DoS attacks. The accuracy and speediness metrics were taken into consideration when conducting our review of the WSN-DS dataset. The J48 method boasts an impressive average processing time of just 0.54 seconds per sample, making it the fastest option available.

In 2020, Kim et al. [16] suggested a CNN-based approach for detecting DoS attacks using the datasets. Many layers were investigated in terms of their ability to do multiclass classification as well as binary classification. Following that, the suggested models were compared to an RNN model, and they outperformed the RNN model in both binary and multiclass classification. Over 97% accuracy was attained by the approach on both classification types.

To provide a stable network and reduce energy consumption in WSN, several related studies have been conducted. To encompass the lifespan of the network by reducing energy usage and reducing transmission delay, this research created the SAD-GHOST approach.

3. PROPOSED METHODOLOGY

This research proposes a novel Selective forwarding Attack Detection utilizing the blue monkey optimized Ghost net (SAD-Ghost) method has been proposed. Initially, Blue Monkey optimization is used for optimal cluster head selection based on node degree and density. Moreover, cluster data are pre-processed using Tokenization, Normalization, and Reduction. The proposed method is utilized to detect intrusion in WSN and classify Normal, DDOS, Grey hole, and Blink litter. Figure 1 shows the suggested method's general block diagram.

3.1 Blue Monkey Optimization using Cluster Head Selection

The blue monkey optimization strategy is utilized to suggest cluster head selection.

3.1.1 The inspiration for the blue monkeys

To add to security, Cercopithecus mitis works along with Cercopithecus ascanius. Males abandon the Cercopithecus mitis social system once they reach adulthood; hence, it is predominantly a female-dominated system. Cercopithecus mitis males rarely engage with the young. Young male Cercopithecus mitis should leave as soon as possible because it is a local species, which will increase their chances of surviving. They challenge the head of another family's family. If they succeed, they will control the family's leadership, which will give the young men the opportunity to socialize and get food and a place to live. The species, Cercopithecus mitis, is nomadic. Due to the accessibility of fruits and basic features, such as grander fruit patches, blue monkeys tend to waste time in forests.

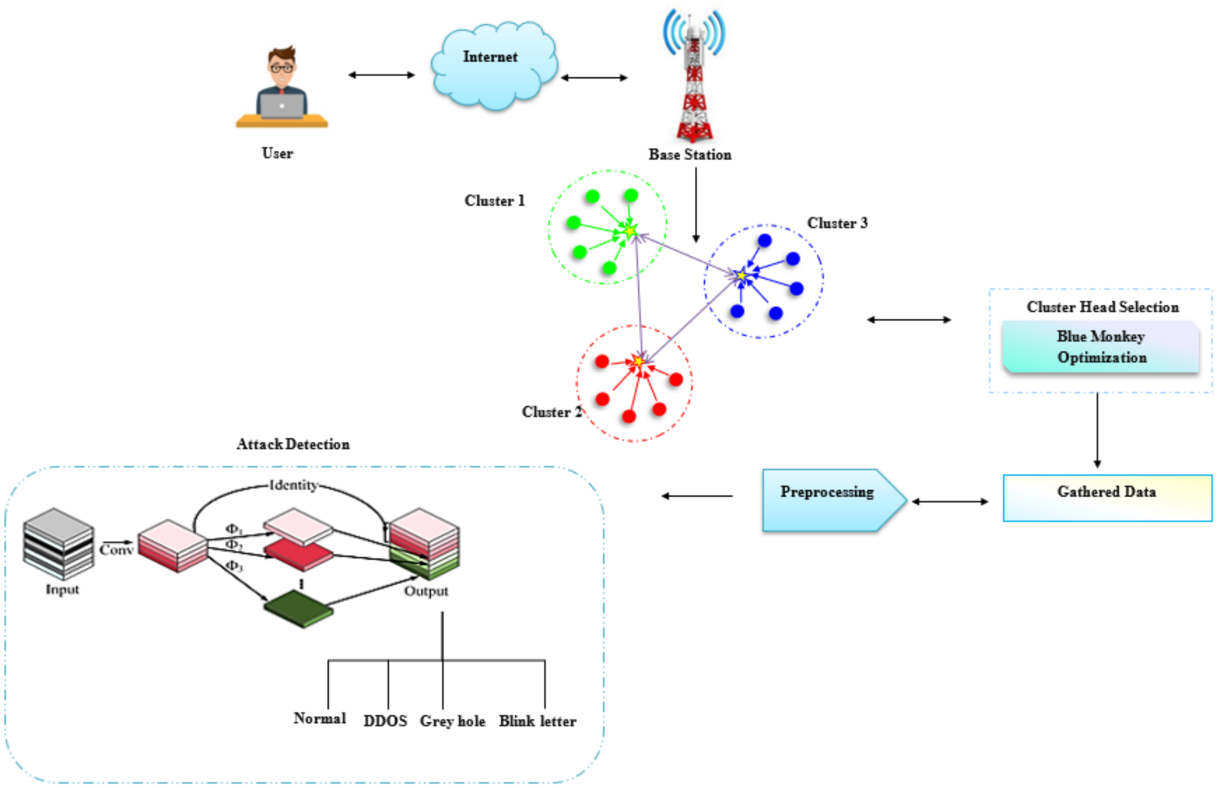


Figure 1. Proposed SAD-GHOST block diagram

Different from other species of monkeys are blue ones. They typically remain with their natal groupings since these social systems are dominated by women. However, after they reach adulthood, males leave their groupings. In the majority of blue monkey groups, there is just one male and many females and young children. This makes inbreeding more difficult. When the males reach adulthood, they leave the group and join another one, although it may take some time for them to find a new group, so they can appear to be lone males. When it comes to social interactions, blue monkeys lack robust instincts. There isn't much time for socializing; it usually happens while people are playing and taking care of one another.

The infants interact with their mothers as well as the other adults in the group. Because of this, these infants are rarely spotted with their male counterparts. Baby handlers handle all operations. The young females carry and protect the newborns as they take care of them. Babies are taught to react to all monkeys through this exercise. This computer program mimics the movements of the blue monkey. The monkey's area unit requires each cluster to move across the search space to mimic these kinds of interactions. Regarding the previous point, in a social setting, the stronger monkey and the one that begins scavenging for food across long distances are not included in the conventional field of view.

Additionally, male Cercopithecus mitis are territorial animals; thus, to get an advantage over the dominant male of a different family, they should leave the nest as soon as possible. The relationship between the males and the younger individuals of the species is currently minimal to nonexistent. If a man defeats another man, he takes over as the head of the family and can provide food, housing, and socializing for young men. Blue monkeys' new position in the group is determined by their previous position. Here is the conceptual framework of BMO equations:

Initialize the population: Generate an initial population of solutions randomly.

$$X_i = (X_{i1}, X_{i2}, \dots \dots \dots X_{id}) \,, \qquad i = \{1,2,\dots.N\} \qquad (1)$$

where N is the population size and d is the dimension of the problem.

Evaluate the Fitness: Calculate the fitness value of each solution based on the objective function.

$$f(X_i), \qquad i=1,2,\dots..N \qquad (2)$$

Update positions: Update the positions of the monkeys based on their foraging behavior.

- **Exploration:** Monkeys explore the search space to find better solutions.

$$X_i^{(t+1)} = X_i^{(t)} + \alpha.rand.(X_j^{(t)} - X_k^{(t)}) \qquad (3)$$

where X_j and X_k are randomly selected monkeys, α is a constant, and a rand is a random number between 0 and 1.

- **Exploitation:** Monkeys exploit the known good solutions to refine their positions.

$$X_i^{(t+1)} = X_i^{(t)} + \beta.rand.(X_{best}^{(t)} - X_i^{(t)}) \quad (4)$$

where $X_{best}^{(t)}$ is the best solution found so far and β is a constant.

Boundary check: Ensure that the new positions are within the predefined boundaries of the search space.

$$X_i^{(t+1)} = \max\left(min\left(X_i^{(t+1)}, X_{max}\right), X_{min}\right) \quad (5)$$

Where X_{max} and X_{min} are the upper and lower boundaries of the search space

Algorithm 1: Blue Monkey Optimization
Initialize the blue monkey and children population BI (I=1...n).
Initialize power rate R and Weight w.
Distribute the fitness of children and all blue monkeys in each group
Calculate the fitness of children and all blue monkeys in each group.
For each group, select the worst value and the best value of fitness and store it in the current best. While children select the best fitness
T=1.
While (T≤ $maximum$ number of iterations)
Swapping the worst fitness in each group by the fitness in the children group.
Update the Rate and x position of all blue monkeys in each group by Equations 1 and 2.
Update Rate and x position of children by Equations 3 and 4.
Update the fitness of all blue monkeys and children.
Update Current Best:
New Best is better than current Best then current Best=New Best.
T=T+1.
End While
Return the optimal blue monkey.

3.2 Pre-Processing of Data

Network traffic processing produces observations represented by feature vectors. These can be classified in to four types they are Normal, DDOS, Grey hole, and blink litter used as input for data mining or machine learning algorithms. As these algorithms can learn from past data, they can automatically classify future observations.

3.2.1 Normalization

To select typical network traffic for testing and training, it is necessary. These datasets must be identified along with information regarding the normalcy or abnormality of the link. It can be difficult and time-consuming to label network traffic.

3.2.2 Tokenization

Its goal is to provide new features that are more discriminative than the original feature set. Machine learning algorithms may greatly benefit from this. Features can be created by hand or with the use of data mining techniques, including frequent-episode mining, association mining, and sequence analysis.

3.2.3 Reduction

It is frequently used to reduce the dataset's dimensionality by removing any superfluous or pointless characteristics. Known as feature selection, this optimization technique is frequently employed to mitigate "the curse of dimensionality. Consider feature extraction as an effective method for reducing data, achieved by transforming the original feature set into a more concise set of new features. PCA, a well-regarded linear technique, is widely used for efficient data reduction.

3.3 Intrusion detection via ghost network

Ghost network is used for the advantage of feature map redundancy and achieves better accuracy and latency than other lightweight networks by striking a balance between accuracy and real-time. ARM-based embedded systems enable exceptional algorithmic performance.

The Ghost Net harnesses the power of the Ghost module, incorporating cost-effective operations and standard convolution techniques. Through this approach, it efficiently merges the M layers of the original feature maps using a single convolution. The process involves two key sections: one that generates the necessary feature concentrations (m) through 1x1 ordinary convolution for identity, and another that utilizes depth-separable convolution blocks to stack and linearly transform the original feature maps (m) into compelling 'Ghost' feature maps layer by layer. These Ghost feature maps are then seamlessly combined with the m feature maps following the identity to create innovative new feature maps.

$$n = m \times s. \quad (6)$$

Considering that a fundamental mapping has been included in the Ghost module and

$$\mathrm{m} \cdot (\mathrm{s}-1) = \mathrm{ns} \cdot (\mathrm{s}-1) \quad (7)$$

Every linear transformation operation should have d × d as its kernel. Theoretically, the Ghost module can accelerate regular convolution by the following ratio:

$$\mathrm{RS} = \frac{\mathrm{n \cdot w' \cdot h' \cdot c \cdot k \cdot k}}{\frac{n}{s} \cdot \mathrm{w' \cdot h' \cdot c \cdot k \cdot k} + (\mathrm{s}-1) \cdot \frac{n}{s} \cdot \mathrm{w' \cdot h' \cdot c \cdot d \cdot d}} = \frac{\mathrm{c \cdot k \cdot k}}{\frac{1}{s} . \mathrm{c \cdot k \cdot k} + (\mathrm{s}-1) \cdot 1\mathrm{s \cdot d \cdot d}} \approx \frac{c.s}{c+s-1} \approx \mathrm{s} \quad (8)$$

The width and height of the yield picture, represented as w' and h', are determined by the number of input channels, denoted as c, received by the convolution kernel. It has improved GhostNet in a few ways since the system in this study is meant to identify a fire in remote-sensing images. More specifically, it enhances the Ghost module with dynamic convolution to improve its ability to adjust to the compound and ever-changing morphology of flames.

The decision to utilize dynamic convolution to enhance GhostNet's Ghost module was influenced by existing literature. Dynamic convolution involves weighting four convolution kernels of the same dimension based on input characteristics.

$$out(x)=\alpha((\partial 1k1+\partial 2k2+\partial 3k3+\partial 4k4)\times x) \quad (9)$$

If each convolution kernel is represented by ki, the convolution operation is denoted by x, the activation function is α, and "the parameter for weighting that depends on the input sample." is αi. Equation 7 shows that αi is the result of focusing on the four essential calculations inside the dashed box.

$$\partial I\ (x)=Sigmoid\ (GAP(x)\ R) \quad (10)$$

where R is the matrix representing the relationship between the number of convolutional kernels and the input dimensions. The firmness of the feature layers to acquire global spatial information is represented by the GAP, and the weights of the four convolution kernels that were formed are represented by the sigmoid function. By integrating the data from several convolutional kernels, dynamic convolution broadens and deepens the network, enhancing the algorithm's ability to extract features. As deep learning progresses, there are an increasing number of various network architectures; nevertheless, as data sizes increased, researchers started focusing on sparse memory, and computational power was leveraged to create compact network models.

3.3.1 Ghost Bottleneck

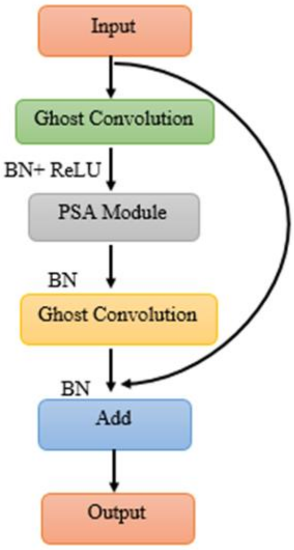


Figure 2. Ghost Bottleneck

Depicts the layout of a bottleneck with two-step sizes. A bottleneck is made by stacking two Ghost modules on top of one another, connecting a residual structure, and then further deepening the network structure. ResNet and the Bottleneck vary in that the Bottleneck adds the same residual structure to the output and adds a deep separable convolution operation with the middle. The goal is to successfully avoid the issues of performance deterioration and gradient vanishing. A batch normalizing procedure is carried out after each layer, and the ReLU function is engaged following the first bottleneck structure. With this structure, the problems of performance degradation and gradient vanishing are successfully avoided. After the first bottleneck structure, each layer is subjected to a batch normalization process before the ReLU function is activated. This topology aims to deepen the network structure while reducing the size of the feature layer.

4. RESULT AND DISCUSSION

To evaluate the performance of the proposed SAD-GHOST approach, we run extensive simulations in this section. Comparisons between the proposed protocol and RSA, RPL, and DOS are made. To compare the suggested technique with past attempts, the simulation time has been set to 1500 seconds.

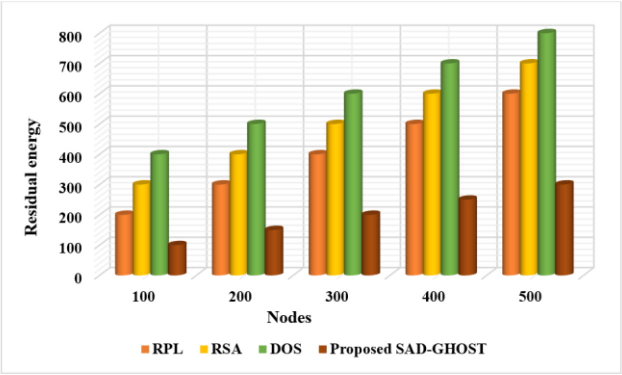


Figure 3. Residual energy vs. number of sensor nodes

Figure 3 illustrates that residual energy diminishes as the number of nodes decreases. Additionally, the proposed SAD-GHOST technique outperforms other routing methods, surpassing previous strategies in terms of average residual energy by 22.15%, 19.45%, and 10.34%.

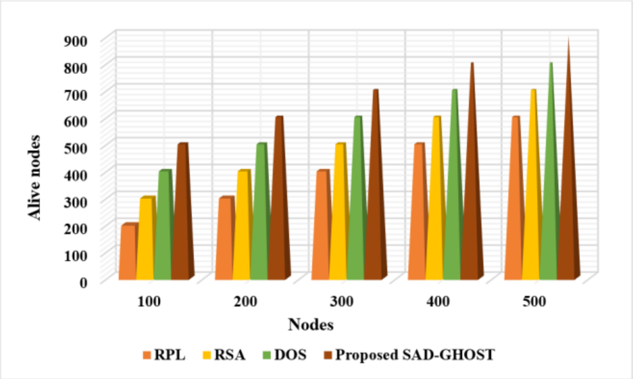


Figure 4. Alive nodes vs. number of sensor nodes

The proposed SAD-GHOST approach generates more alive nodes than the earlier DSO, RSA, and RPL methods, as demonstrated in Figure 4. The SAD-GHOST strategy is found to be superior when compared to older techniques, the proportion of active network nodes.

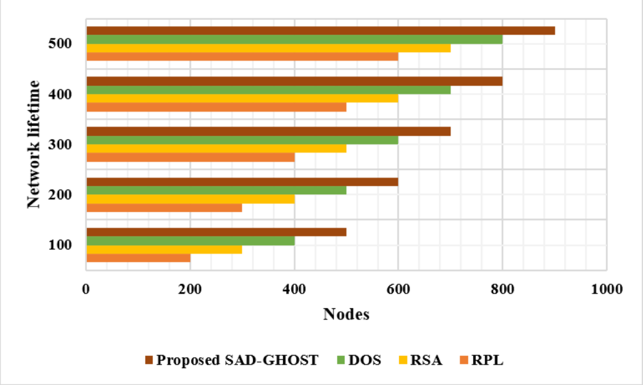


Figure 5. Network lifetime vs. number of Sensor nodes

Figure 5 shows the lifespan of the network for different node counts. The analysis demonstrates that even with a large number of nodes, the suggested SAD-GHOST model lengthens the network's lifespan. The comparable network lifespan determined by the existing methods (RPL, RSA, and DOS) is planned at 22.34%, 20.13%, 17.23%, and 19.04%, respectively.

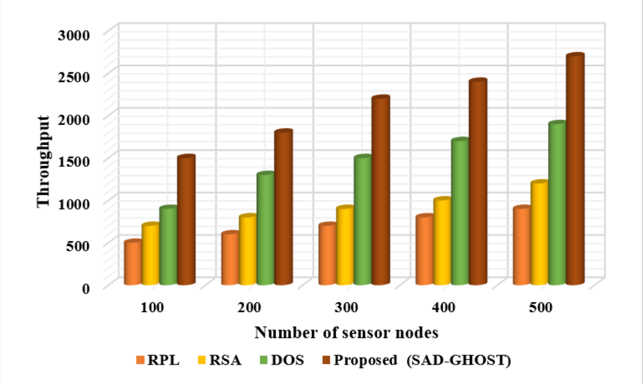


Figure 6. Throughput Vs Number of Sensor nodes

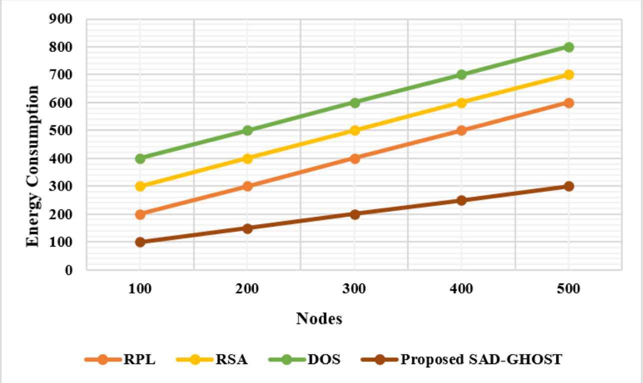


Figure 7. Energy Consumption Vs Number of Sensor Nodes

Figure 6 presents a throughput comparison between the SAD-GHOST strategy and previous methods. The results demonstrate that the proposed method performs exceptionally well in terms of throughput for all sensor nodes. One of the most crucial factors in the development of QoS is throughput. Thus, network scalability is guaranteed. Attaining optimal network performance is not solely dependent on network lifespan. Consequently, QoS factors are crucial for enhancing network efficiency and bolstering the dependability of routing algorithms. Any network's routing protocol's scalability is directly correlated with its throughput. Here, the proposed method's scalability reaches a limit when the number of nodes rises above 1000.

Figure 7 provides a detailed comparison of the SAD-GHOST technique to existing methods (RPL, RSA, and DOS), in terms of energy consumption vs nodes. The figure illustrated the RPL protocol's incapability to outperform other strategies with a maximum amount of energy usage. Additionally, compared to the RSA method, DOS obtained a somewhat reduced energy use. The proposed SAD-GHOST methodology, however, surpassed existing strategies with a lower energy consumption, while the present approach achieved a slightly higher energy consumption.

5. CONCLUSION

In this research proposed a novel selective forwarding attack detection method based blue monkey-optimized ghost net (SAD-Ghost) method. To identify network threats, Blue Monkey optimization based on the hazard model is built in this case. A proposed technique to improve detection accuracy and minimize computation. The primary goal of the research is to develop a Selective Forwarding Attack Detection utilizing a blue Monkey optimized Ghost net to improve network lifetime. Initially, Blue Monkey optimization is used for optimal cluster head selection based on node degree and density. Moreover, cluster data are pre-processed using Tokenization, Normalization, and Reduction. The proposed method is utilized to detect intrusion in WSN and classify Normal, DDOS, Grey hole, and Blink litter. The experimental analysis demonstrates that the proposed method achieves a packet delivery rate of 97.6%, 95.3% and 90.50% and reduces energy consumption by 19.6%, 12.5% and 17.4% compared to existing clustering-based routing methods. Consequently, the proposed technique surpasses current methods in terms of network lifetime, energy efficiency, and packet delivery performance. As a result, the proposed techniques outperform existing methods based on network lifetime, alive nodes, and residual energy.

CONFLICTS OF INTEREST

This paper has no conflict of interest for publishing.

FUNDING STATEMENT

Not applicable.

ACKNOWLEDGEMENTS

The author would like to express his heartfelt gratitude to the supervisor for his guidance and unwavering support during this research for his guidance and support.

REFERENCES

[1] M. Faris, M.N. Mahmud, M.F.M. Salleh, and A. Alnoor, "Wireless sensor network security: A recent review based on state-of-the-art works", *International Journal of Engineering Business Management*, vol. 15, pp.18479790231157220, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[2] M. Krichen, M.S. Abdalzaher, M. Elwekeil, and M.M. Fouda, Managing natural disasters: An analysis of technological advancements, opportunities, and challenges. *Internet of Things and Cyber-Physical Systems.* 2023. [CrossRef] [Google Scholar] [Publisher Link]

[3] R. Yarinezhad, and S.N. Hashemi, A sensor deployment approach for target coverage problem in wireless sensor networks. *Journal of Ambient Intelligence and Humanized Computing,* vol. 14, no. 5, pp.5941-5956, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[4] T. Jabeen, I. Jabeen, H. Ashraf, N.Z. Jhanjhi, A. Yassine, and M.S. Hossain, An intelligent healthcare system using IoT in wireless sensor network. *Sensors*, vol. 23, no. 11, pp.5055, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[5] P.A.D.S.N. Wijesekara, and S. Gunawardena, A Review of blockchain technology in knowledge-defined networking, its application, benefits, and challenges. *Network*, vol. 3, no. 3, pp.343-421, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[6] B. Kapoor, B. Nagpal, and M. Alharbi, Secured healthcare monitoring for remote patient using energy-efficient IoT sensors. *Computers and Electrical Engineering,* vol. 106, pp.108585, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[7] A. Morchid, R. El Alami, A.A. Raezah, and Y. Sabbar, Applications of internet of things (IoT) and sensors technology to increase food security and agricultural Sustainability: Benefits and challenges. *Ain Shams Engineering Journal*, pp.102509, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[8] F. Zijie, M.A. Al-Shareeda, M.A. Saare, S. Manickam, and S. Karuppayah, "Wireless sensor networks in the internet of things: review techniques, challenges, and future directions," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 31, no. 2, pp.1190-1200, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[9] A. Suhag, and A. Daniel, Study of statistical techniques and artificial intelligence methods in distributed denial of service (DDOS) assault and defense. *Journal of Cyber Security Technology,* vol. 7, no. 1, pp.21-51, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[10] Ö. Aslan, S.S. Aktuğ, M. Ozkan-Okay, A.A. Yilmaz, and E. Akin, A comprehensive review of cyber security vulnerabilities, threats, attacks, and solutions. *Electronics*, vol. 12, no. 6, pp.1333, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[11] M. Premkumar, and T. Sundararajan, "Dldm: Deep learning-based defense mechanism for denial-of-service attacks in wireless sensor networks", *Microprocess Microsyst.* Vol. 79, pp. 103278, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[12] M. Asad, M. Asim, T. Javed, M.O. Beg, H. Mujtaba, and S. Abbas, "Deepdetect: detection of distributed denial of service attacks using deep learning", *Comput J.* vol. 63, no. 7, pp. 983–94. 2020. [CrossRef] [Google Scholar] [Publisher Link]

[13] S. Salmi, and L.Oughdir, "Cnn-lstm based approach for dos attacks detection in wireless sensor networks". *Int J Adv Comput Sci Appl. vol*. 13, no. 4, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[14] R, Wazirali, and R. Ahmad, "Machine learning approaches to detect dos and their effect on wsns lifetime", *CMC-Comput Mat Contin.* Vol. 70, no. 3, pp. 4921–46, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[15] J. Kim, H. Kim, M. Shim, and E. Choi, "Cnn-based network intrusion detection against denial-of-service attacks", *Electronics.* [CrossRef] [Google Scholar] [Publisher Link]

[16] L. Alsulaiman, and S. Al-Ahmadi, "Performance evaluation of machine learning techniques for DOS detection in wireless sensor network", arXiv preprint arXiv:2104.01963. 2021. [CrossRef] [Google Scholar] [Publisher Link]

AUTHORS

Jeyaraman Sathiamoorthy is currently working as a Professor RMK ENGINEERING COLLEGE, Kavarapettai in Chennai. He has completed M. Tech (CIT) and Ph. D from Manonmaniam Sundaranar University, Tirunelveli. He has 20 years of teaching experience and she has published papers in Network Security in National and International journals and has presented in International Conferences and Seminars. His current area of interest is Programming Languages, Algorithms and ad-hoc networks especially MANET, VANET, FANET and Underwater Communication.

M. Usha is currently working as a Professor cum Assistant Director in MEASI Institute of Information Teahnology, Chennai. She has completed M.C.A. and M. Phil in Computer Science from Bharathidasan University, Trichy. She has also done her M. Tech (CIT) and Ph. D from Manonmaniam Sundaranar University, Tirunelveli. She has 20 years of teaching experience and she has published papers in Network Security in National and International journals and has presented in International Conferences and Seminars. Her current area of interest is Operating Systems, Algorithms and ad-hoc networks especially MANET, VANET, FANET and Underwater Communication.

P. Senthilraja is currently working as an Assistant Professor at BV Raju Institute of technology, Narsapur in Telengana. He has completed M.Tech(CIT) from Manonmaniam Sundaranar University, Tirunelveli. He has 12 years of teaching experience and he has published papers in Network Security in National and International journals. His current area of interest is Machine learning, network security, ad-hoc networks especially MANET and VANET.

Arrived: 20.04.2024

Accepted: 08.06.2024

International Journal of Data Science and Artificial Intelligence (IJDSAI)
Volume 02, Issue 03, May – June (2024)

RESEARCH ARTICLE

IOT-CENTRIC DATA PROTECTION USING DEEP LEARNING TECHNIQUE FOR PRESERVING SECURITY AND PRIVACY IN CLOUD

C. Senthil Singh [1, *], Sameena Naaz [2] and G. Saranya [3]

[1] Department of Electronics and Communication Engineering, Shadan Women College of Engineering and Technology, Hyderabad, India.
[2] Department of Computer Science, Roehampton University, London, SW15 5PH, United Kingdom.
[3] Department of computer science and Engineering, SA Engineering college, Poonamallee, Tamil Nadu 600077 India.

[*]*Corresponding e-mail: senthilsingh@gmail.com*

Abstract – The Internet of Things (IoT) describes a system where interconnected physical objects are connected online. As the collection and sharing of vast amounts of personal data grow, so do concerns over user privacy within IoT environments. While IoT devices offer significant advantages in terms of productivity, accuracy, and financial benefits by minimizing human intervention and providing exceptional flexibility and convenience, they also face challenges related to communication overhead, security, and privacy. To address these issues, a novel Internet of Things-based Cloud Information Security Preservation (IoT-CISP) has been proposed. This approach enhances the model's effectiveness and ensures security by first separating sensitive data from non-sensitive data using an SVM classifier, and then employing this data for partial decryption and analysis. Sensitive data is protected through Okamoto-Uchiyama encryption, ensuring that data storage, analysis, and sharing are conducted securely to maintain the system's safety and privacy. The effectiveness of this novel method was assessed against existing methodologies using parameters like precision, accuracy, F1 score, and recall, revealing its superior security and efficiency compared to other schemes. Results demonstrate that the IoT-CISP approach offers encryption times that are 31.24%, 23.12%, and 33.03% shorter than those of the CP-ABE, GDBR, and HP-CPABE algorithms, respectively.

Keywords *– cloud computing, Internet of things, Support vector machine, Okamoto Uchiyama.*

1. INTRODUCTION

IoT devices, generate enormous amounts of data but lack the storage and the processing capability to process it, which are the foundation of the cyber-physical system [1, 2]. The cloud platform has established itself as a top-tier method of storing, analyzing, and exchanging data with numerous stakeholders for the best results [3, 4]. However, it is not advisable to put your faith in a third-party cloud platform, particularly for preserving private information, as renting data to the cloud results in the gadget losing ownership of it [5].

An average data loss expense worldwide will be $4.35 million in 2022, up to 2.6% and 12.7% from 2021 and 2020, respectively, according to research by IBM and Ponemon Institute [6]. These factors have made data protection a major problem, which has inspired researchers to provide strategies for maintaining data privacy [7]. Most of the models that are now in use are based on encryption techniques [8], and differential privacy, but they have a limit According to research by IBM and the Ponemon Institute, the average cost of a data breach globally will be $4.35 million in 2022, an increase of 2.6% and 12.7% from 2021 and 2020, respectively.

These reasons have made data protection a significant issue, prompting researchers to offer techniques for preserving data privacy. The majority of models now in use are based on encryption methods as well as differential privacy, although their accuracy, utility, and efficiency might be improved. As far as the authors are aware, there are currently no models that adequately balance the accuracy and privacy of the data that is being outsourced [9,10]. To overcome these challenges Internet of Things based Cloud Information Security preservation (IoT-CISP) has been proposed. The following are the paper's main contributions:

- The proposed IoT-CISP approach enhances the model's effectiveness and ensures security by first separating sensitive data from non-sensitive data using an SVM classifier, and then employing this data for partial decryption and analysis.
- Sensitive data is protected through Okamoto-Uchiyama encryption, ensuring that data storage, analysis, and sharing are conducted securely to maintain the system's safety and privacy.

- The effectiveness of this novel method was assessed against existing methodologies using parameters like precision, accuracy, F1 score, and recall, revealing its superior security and efficiency compared to other schemes.

The remaining sections are arranged as follows: Section 3 presents the proposed framework. The experimental setting is demonstrated in Section 4. The collected results are then detailed in Section 5. With a conclusion and future work, the paper is completed.

2. LITERATURE SURVEY

In 2021, Zhang et al. [11] proposed the HP-CP-ABE hidden access policy to safeguard data security and authenticate authorized users. One potential issue is the possibility of attackers launching attribute values to reveal. The parameter data is included in several HP-CP-ABE methods' access controls, employing guessing attacks (AVGA). To address the computational PBDHE assumption as a means to establish the selective IND-AVGA security of the proposed scheme, this is the first time this assumption has been applied.

In 2023 Vaidya S [12] Suggested a hidden approach framework paired with a based encrypted approach to improve security associated with healthcare IoT information. The system ensures precise control over access to encrypted data and safeguards the privacy of clinical clients. The proposed system far surpasses existing systems in terms of, security, storage load, and computing efficiency, especially when the approach structure is concealed. It presents a new approach for strongly transmitting data in the context of IoT.

In 2023 Wang. C et al. [13] Proposed a robust data encryption scheme called Attribute Hiding and Multiple Authorization Centers-based Data Hierarchical Encryption Scheme (AH-MAC-DHE). This strategy Maintains private data by disguised access controls and user features. to handle the problem of private data being leaked. Assuming judgmental q-parallel the coefficient of the Bilinear Diffie-Hellman, they have shown that AH-MAC-DHE is reliable and offers security for privacy as well as anti-collusion. According to experimental findings, AH-MAC-DHE performs better than current methods.

In 2022 Li, M., Xiao, D., et al. [14] proposed based on compressive sensing, utilizing private cloud for three different levels of cloud service users. This enables the sensor-cloud system to offer a variety of multimedia service levels and security assurances. From the standpoint of consumers of cloud-based services, ensuring the privacy of essential data is a challenge. The suggested approach successfully balances the relationships between cloud service providers, sensor network suppliers, and cloud service customers, according to theoretical studies and experimental simulations.

In 2022 Wei et al. [15], proposed Scourge modeling techniques for addressing privacy of data threats in independent systems, and we have analyzed these techniques in the context of GDPR compliance. Additionally, discussed the challenges and identified gaps, offering suggestions for a new modeling technique. This technique not only models' conventional risks to data privacy but also efficiently does GDPR compliance checks.

In 2023 Huang, B. et al [16] Proposed Ciphertext-Policy Attribute-Based Encryption (CP-ABE) as a lattice-based encryption technique. Compared to techniques built on the Learning with Error problem, this particular approach is more efficient because it is based on the Ring Learning By Error issue. Due to unreliable cloud service providers, the approach addresses security and access control concerns for critical data. Evaluations and experimental simulations demonstrate the scheme's excellent applicability and efficiency.

3. PROPOSED METHODOLOGY

A malicious utility provider might take data that has been outsourced from the cloud, store it, analyze it, and share it with the parties involved to gather private information that could be abused. As a result, protecting data has grown to be a difficult undertaking that must be handled carefully. This paper provides a safe data protection technique for preserving confidentiality in a cloud context to address this crucial and difficult problem. It does so by effectively separating the data into sensitive and non-sensitive categories using an SVM classifier, partially decrypting the data, and performing data analysis that increases the model's effectiveness while maintaining security. Okamoto Uchiyama encryption has been used to protect sensitive data. By carrying out safe data storage, analysis, and exchange, the model guarantees the security and privacy of the system. Specific criteria, including precision, accuracy, F1-score, and recall, have been used to compare the suggested method to the existing methodologies. Figure 1 shows the overall framework for the suggested work.

3.1 Support Vector Machine

Support vector machines (SVMs) are used to collect data from IoT devices, which is divided into two categories: sensitive and nonsensitive data. SVMs are a type of supervised machine learning technology that converts complex, highly nonlinear circumstances into binary classification models [5]. The SVM must construct the decision surface, which is a hyperplane, utilizing data samples to maximize the margin around it. During the training stage of the algorithm, each data sample is assigned a class designation and the projected value. The data sample contains what are referred to called characteristics, these are the variables in the data that specify the data sampling vector's activity. The weights applied to each input feature and a collection of support vectors that construct the ideal hyperplane are used to forecast the results of the SVM training phase. In contrast to other neural networks, the SVM maximizes the number of nonzero weights while lowering the overall number of nonzero weights by maximizing the margin. These only match the important traits that provide information that is useful for selecting the hyperplane.

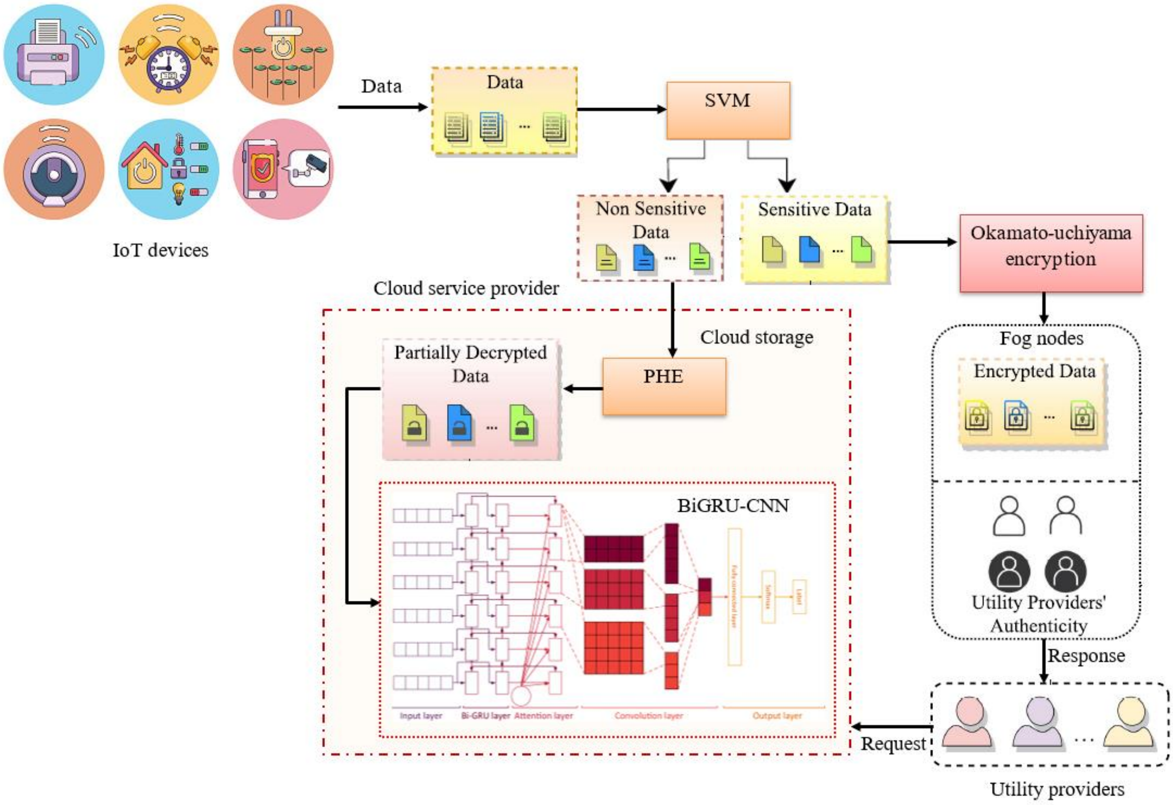


Figure 1. Block diagram of the proposed method

The kernel function modifies the data dimensions to define the hyperplane's form, which is a crucial step in the SVM. Simply put, the kernel function increases the hyperplane's size to help distinguish between the classes. The employment of many kernel types is possible, including the polynomial, linear, both sigmoid and the Gaussian radial basis function. The type of data sample affects how each kernel performs. The simplest kernel, the linear kernel, performs better when applied to linear problems. The supplied characteristics are combined by the polynomial, RBF, and sigmoid kernels to produce support vectors. They work best with non-linear data, but their complexity depends on how many additional features they find. In Figure 2, the SVM flowchart is displayed.

3.2 Okamoto Uchiyama Encryption

Numerous concepts from number theory, discrete mathematics, and abstract algebra are used in the Okamoto-Uchiyama cryptosystem. Numerous of these ideas are fundamental and are applied in various areas of cryptography. However, even though beyond integer computations, there is no need to provide detailed or rigorous treatments, those essential notions are still important. They are not sufficiently covered in mathematics curricula in underdeveloped and developing countries, by examining the fundamental concepts and mathematics used in the Okamoto-Uchiyama algorithm.

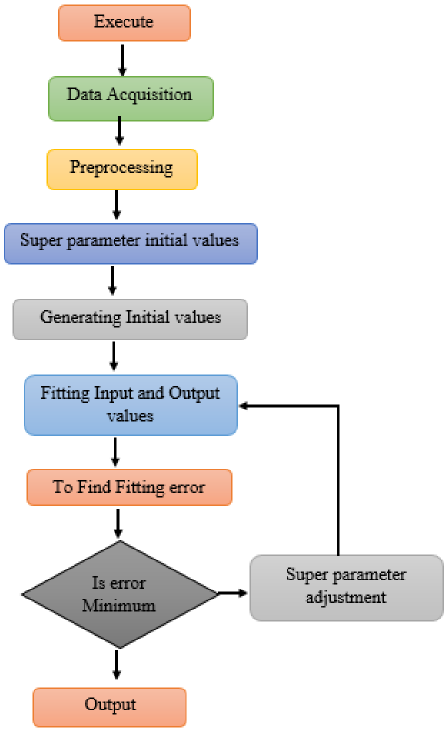

Execute
Data Acquisition
Preprocessing
Super parameter initial values
Generating Initial values
Fitting Input and Output values
To Find Fitting error
Is error Minimum
Super parameter adjustment
Output

Figure 2. Flow chart of Support Vector Machine (SVM)

3.2.1 Key Generation

The following procedure generates a public/private key:

1. Create the two big primes, A and B.
2. Compute N=B^2A
3. Select an integer with a random value G2 n-1 such that G(B-1)
4. Computes H=GN mod N.

Next, we have (N, G, H) as the public key and (B, A) as the private key.

3.2.2 Encryption

Using the public key (B, A), the following can be done to encrypt a message M < B.

1. A random number R1n-1 should be chosen so that G(B-1)
2. Compute C= GN HR mod N.

The value C is the encryption of M

3.2.3 Decryption

An encrypted message C can be decrypted with the private key (B, A) as follows:

1. Compute P= l (C^{B-1}mod B^2
2. Compute Q= l (G^{B-1}mod B^2 p and Q will be integers.
3. Using the extended Euclidean technique, compute the opposite value of Q modulo B.
   $Q' = Q^{-1}\ mod\ B$
4. Compute M=$PQ'mod\ B$

The value M is the decryption of C.

3.3 Partially Homomorphic Encryption Using Sensitive Data

Schemes for partial homomorphic encryption (PHE) provide the performance of certain accurate processes on scrambled data. The PHE cryptosystem uses public keys. It comprises three stages: encryption, decryption, and key creation. Specifically, since it possesses the public key and secret key throughout the phase of generating keys, every device receives the public key, which is used for encrypted data. The system operator or just one agent has access to the private key needed to decrypt messages. Specifically, there is a time variation in both the public and private keys.

$$Enc(L_1)^{\circ}Enc(L_2) = Enc(L_1 + L_2) \tag{1}$$

In particular the ciphertext of $L_1 + L_2)$ can be obtained from the ciphertexts of L_1 and L_2 As seen in Algorithm 1, the three roles execute all three PHE encryption system features.

3.4 BiGRU-CNN

Figure 3 shows the process of entering current load data into the CNN network's computational layer from the data source layer. The convolutional layer is used to record data correlation. By applying pooling operations to reduce data dimensionality, the pooling layer improves the efficiency of network learning.

Algorithm 1: PHE Cryptosystem

```
Function Keygen ()
    Output: Public key $K_P$:(N, G), private key:($⋌, \mu$)
    Choose two large prime numbers P and Q of equal bit-length and compute N=P.Q;
    G←N+1;
    $⋌$=$\emptyset(N) = (P-1).(Q-1),$ where $\emptyset(\cdot)$ denotes the Euler's totient function;
    $\mu = \emptyset^{-1}$(N)mod N, which is the modular multiplicative inverse of $\emptyset(N)$;
Function $\mathcal{E}(L)$
    Output: Ciphertext c
    the random $r \in \emptyset^*{}_{-1}$={$z|z \in Z, 0 \leq z < N, gcd(z,n) = 1$};
    element the ciphertext by c=$G^L.r^N$mod $N^2$,where m∈ $Z_N$= {z|z ∈ $Z, 0 \leq z < N\}, c \in Z_N{}^*$2;
Function D(C)
    Output: Message L
    Define the integer division function M ($\mu$)=$\frac{\mu-1}{N}$;
    Generate the plaintext as L=m ($c^{⋌}mod\ N^2$ ). $mod$N;
```

Following data entry into the BiGRU network, the processed data undergoes full learning, which enhances the accuracy of temporal feature extraction even further. The completely connected layer ultimately produces the final forecasting results.

3.4.1 Convolutional Neural networks

The deep design and integration of convolution processing define convolutional neural networks (CNNs), a subclass of feed-forward neural networks. They usually deal with spatial data loss, inefficiencies, and overfitting. Combining both convolutional and pooling layers in the CNN model framework enables efficient feature-learning and classification activities by automatically obtaining features at different scales and levels.

3.4.2 Gated Recruitment Unit (GRU)

One model that processes sequence data using the deep learning technique is called BIGRU. It is predicated on enhancing GRU with the addition of a bidirectional loop structure, which improves the collection of contextual information in sequence data. Two gated loop units are included in every BiGRU unit; one is used to process sequence data in the direction that is forward, and the other is used to process data in the opposite direction.

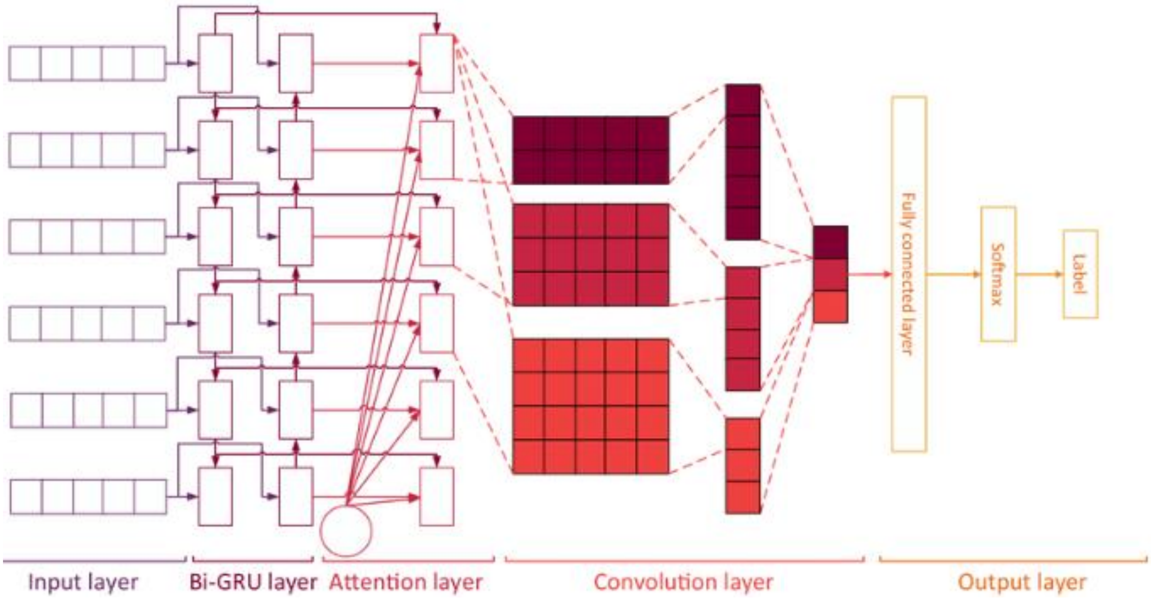


Figure 3. Bi-GRU CNN Neural networks

It lowers the model's training complexity by combining an LSTM's input and remembering gates into just one updating gate complexity and convergence time, as well as its number of parameters, and achieving faster training convergence.

$$r_t = \sigma(W_r x_t + U_r h_{t-1}) \quad (1)$$

$$z_t = \sigma(W_r x_t + U_r h_{t-1}) \quad (2)$$

$$\breve{h}_t = \tanh\,(W x_t + U(\boldsymbol{r_t} \odot h_{t-1})) \quad (3)$$

$$h_t = (1 - z_t)h_{t-1} + z_t \breve{h}_t \quad (4)$$

Here, Wt, Wz, W, Uk, U are the GRU weight matrix values. σ indicates the reasonable sigmoidal function; $tanh$ indicates the tanh function; $\odot$ indicates the element multiplication operation; zt indicates the update gate, which determines the degree of informing of the GRU unit's activation value based on the state of the input and the state of the earlier hidden layer in tandem; rt indicates the rearrange gate, whose informing procedure is comparable to that of zt; The candidate hidden layer I indicated by $h\tilde{t}$, while the hidden layer indicated by ht .

4. RESULT AND DISCUSSION

As shown in the next sections, we tried several experiments in this work to address the privacy issue using deep learning algorithms. By carrying out safe data storage, analysis, and exchange, the model guarantees the security and privacy of the system. Specific criteria, including precision, accuracy, F1-score, and recall, have been used to compare the suggested method to the existing methodologies.

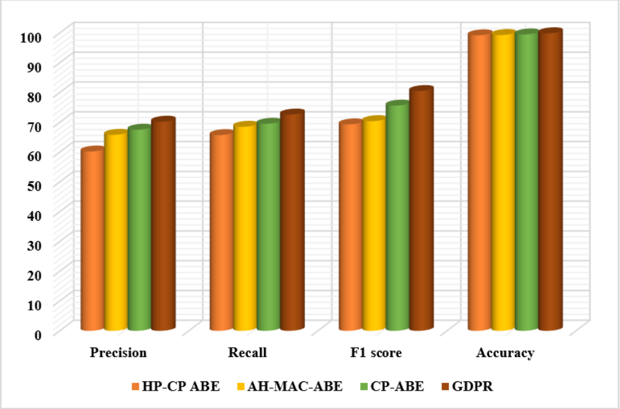


Figure 4. Performance Evaluation

Figure 4 illustrates the findings, which demonstrate that all deep learning methods offer high evaluation metrics in contrast, the averages for benign are 99.92%, 98.85%, and 99.90% for precision, recall, and F1-score.

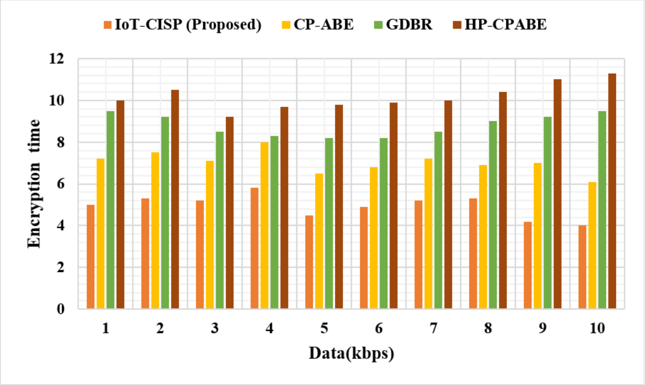


Figure 5. Encryption Time

Figure 5 illustrates the encryption time for the IoT-CISP technique, defined as the duration an encryption algorithm takes to generate ciphertext from plaintext. The encryption time depends on the number of images used in the process.

The proposed method demonstrates faster encryption compared to existing techniques like CP-ABE, GDBR, and HP-CPABE, which require more time. The proposed IoT-CISP technique not only achieves quicker encryption but also performs better with larger datasets. According to the results, IoT-CISP achieves encryption times that are 31.24%, 23.12%, and 33.03% faster than CP-ABE, GDBR, and HP-CPABE, respectively.

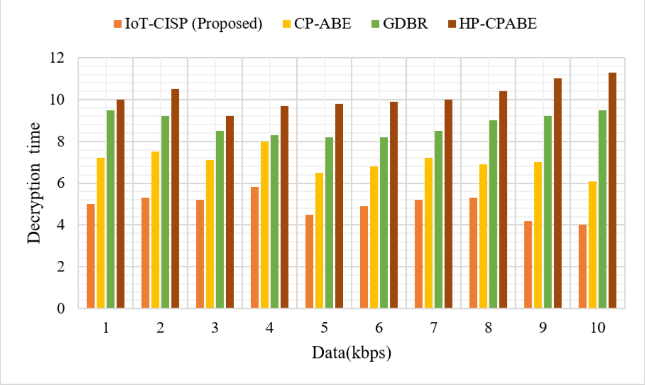


Figure 6. Decryption time

Figure 6 shows the decryption time for the IoT-CISP technique, which refers to the time it takes for the decryption algorithm to convert ciphertext back into plaintext. This decryption time varies based on the number of data involved in the process. The proposed IoT-CISP method achieves faster decryption compared to existing techniques like CP-ABE, GDBR, and HP-CPABE, which require more time. Furthermore, the IoT-CISP technique performs better with larger datasets. The results indicate that IoT-CISP is 21.24%, 13.12%, and 23.03% faster in decryption times compared to CP-ABE, GDBR, and HP-CPABE, respectively.

5. CONCLUSION

In this research a novel Internet of Things based Cloud Information Security Preservation (IoT-CISP) has been proposed. This approach enhances the model's effectiveness and ensures security by first separating sensitive data from non-sensitive data using an SVM classifier, and then employing this data for partial decryption and analysis. Sensitive data is protected through Okamoto-Uchiyama encryption, ensuring that data storage, analysis, and sharing are conducted securely to maintain the system's safety and privacy. The effectiveness of this novel method was assessed against existing methodologies using parameters like precision, accuracy, F1 score, and recall, revealing its superior security and efficiency compared to other schemes. Results demonstrate that the IoT-CISP approach offers encryption times that are 31.24%, 23.12%, and 33.03% shorter than those of the CP-ABE, GDBR, and HP-CPABE algorithms, respectively. In the future, the current model will be tested on additional benchmark datasets as part of our research plans.

CONFLICTS OF INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

FUNDING STATEMENT

Not applicable.

ACKNOWLEDGEMENTS

The author would like to express his heartfelt gratitude to the supervisor for his guidance and unwavering support during this research for his guidance and support.

REFERENCES

[1] K. Singh and I. Gupta, "Online information leaker identification scheme for secure data sharing", *Multimedia Tools Appl.*, vol. 79, no. 41, pp. 31165–31182, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[2] I. Gupta, A. K. Singh, C. N. Lee, and R. Buyya, "Secure data storage and sharing techniques for data protection in cloud environments: A systematic review, analysis, and future directions", *IEEE Access*, vol. 10, pp. 71247–71277, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[3] D. Saxena, I. Gupta, A. K. Singh, and C. N. Lee, "A fault-tolerant elastic resource management framework toward high availability of cloud services", *IEEE Trans. Netw. Service Manage.*, vol. 19, no. 3, pp. 3048–3061, Sep. 2022. [CrossRef] [Google Scholar] [Publisher Link]

[4] R. Gupta, D. Saxena, I. Gupta, A. Makkar, and A. K. Singh, "Quantum machine learning driven malicious user prediction for cloud network communications", *IEEE Netw. Lett.*, [CrossRef] [Google Scholar] [Publisher Link]

[5] I. Gupta and A. K. Singh, "SELI: Statistical evaluation-based leaker identification stochastic scheme for secure data sharing", *IET Commun.*, vol. 14, pp. 3607–3618, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[6] IBM Ponemon, "Cost of a data breach study", 2022. [Online] Available: https://www.ibm.com/security/data-breach[CrossRef] [Google Scholar] [Publisher Link]

[7] I. Gupta and A. K. Singh, "Dynamic threshold-based information leaker identification scheme", *Inf. Process. Lett.*, vol. 147, pp. 69–73, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[8] X.Ma, J.Ma, H. Li, Q. Jiang, and S. Gao, "PDLM: Privacy-preserving deep learning model on cloud with multiple keys", *IEEE Trans. Serv. Comput.*, vol. 14, no. 4, pp. 1251–1263, Jul./Aug. 2021. [CrossRef] [Google Scholar] [Publisher Link]

[9] W. Zhao, W. Yang, H. Wang, T. Zhang, D. Man, T. Liu, J. Lv, and M. Guizani, "Privacy-Preserving Outsourcing of K-Means Clustering for Cloud-Device Collaborative Computing in Space-Air-Ground Integrated IoT", *IEEE Internet of Things Journal*, vol. *10*, no. 23, pp.20396-20407, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[10] S. Fugkeaw, L. Wirz, and L. Hak, "Secure and Lightweight Blockchain-Enabled Access Control for Fog-Assisted IoT Cloud Based Electronic Medical Records Sharing", *IEEE Access*, vol. *11*, pp.62998-63012, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[11] Z. Zhang, W. Zhang, and Z. Qin, "A partially hidden policy CP-ABE scheme against attribute values guessing attacks with online privacy-protective decryption testing in IoT-assisted cloud computing", *Future Generation Computer Systems*, vol. 123, pp. 181-195, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[12] S., Vaidya, A., Suri, V., Batla, I. Keshta, S.S.M. Ajibade, and G. Safarov, "A computer-aided feature-based encryption

model with concealed access structure for medical Internet of Things", *Decision Analytics Journal*, vol. 7, pp. 100257, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[13] C. Wang, J. Pan, J. Lu, and Z. Wu, "A Data Hierarchical Encryption Scheme Based on Attribute Hiding under Multiple Authorization Centers", *Electronics*, vol. 13, no. 1, pp. 125, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[14] M. Li, D. Xiao, H. Huang, and B. Zhang, "Multi-level video quality services and security guarantees based on compressive sensing in sensor-cloud systems", *Journal of Network and Computer Applications*, vol. 205, pp.103456, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[15] N. Azam, L. Michala, S. Ansari, and N.B. Truong, "Data privacy threat modeling for autonomous systems: A survey from the GDPR's perspective", *IEEE Transactions on Big Data*, vol. 9, no. 2, pp.388-414, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[16] B. Huang, J. Gao, and X. Li, "Efficient lattice-based revocable attribute-based encryption against decryption key exposure for cloud file sharing", *Journal of Cloud Computing*, vol. 12, no 1, pp.37, 2023. [CrossRef] [Google Scholar] [Publisher Link]

AUTHORS

C. Senthil Singh received his B. E, MTech, PhD degree in Information and Communication from Anna University, Chennai, India. He has a great flair for teaching and research and has a total experience of 22 years in teaching in Engineering. His professional interests include VLSI, Wireless Communication, Embedded Systems and Telemedicine.

Sameena Naaz is currently working as a Senior Lecturer at the University of Roehampton, London, United Kingdom. She holds the post of Professor at the Department of Computer Science and Engineering, Jamia Hamdard, New Delhi, India (On leave). She has a total experience of more than 23 years. She received her B.Tech (Computer Engg.) from Aligarh Muslim University, in 1998 and the M.Tech. Degree in Electronics with Specialization in Communication and Information Systems from Aligarh Muslim University, in 2000. She completed her Ph. D from Jamia Hamdard in the field of distributed systems in year 2014. Sameena Naaz has published several research articles in reputed International Journals and Proceedings of reputed international conferences published by IEEE and Springer. Her research interests include Distributed Systems, Cloud Computing, Big Data, Machine Learning, Data Mining and Image Processing.

G Saranya received her B.E degree in Computer Science and Engineering from Anna University, Chennai and M.E degree in Computer Science and Engineering from Hindustan University, Chennai. She started her career as an Assistant Professor and has 9 years and 6 months of experience. Currently she is working as an Assistant Professor in S.A. Engineering College, Chennai. Her research interests include Deep Learning and Cloud Computing. She is a lifetime member of ISTE.

Arrived: 29.04.2024

Accepted: 12.06.2024

International Journal of Data Science and Artificial Intelligence (IJDSAI)
Volume 02, Issue 03, May – June (2024)

RESEARCH ARTICLE

BLOCK CHAIN ENABLED DATA SECURITY USING BLOWFISH ALGORITHM IN SMART GRID NETWORK

R.C. Ilambirai [1, *], S. Lourdu Jame [2] and P.U. Poornima [3]

[1] Department of Electrical and Electronics Engineering, SRMIST, Kattankulathur, Chennai, Tamil Nadu, 603203 India.
[2] Department of Electrical and Electronics Engineering, SRMIST, Kattankulathur, Chennai, Tamil Nadu, 603203 India.
[3] Department of Electrical and Electronics Engineering, SRMIST, Kattankulathur, Chennai, Tamil Nadu, 603203 India.

[*]*Corresponding e-mail: ilambirai12@gmail.com*

Abstract – Smart Grid provides a reliable and efficient end-to-end delivery system. Data on each user's unique electricity consumption is given in real time. It also enables utilities to control and monitor the electrical system in real-time, helping them to reduce power outages. Users' privacy is a significant issue in addition to the usual security issues. Data about power usage may be used to infer private information about users by entities with access to the data. To solve these problems, a Block-Chain-based Secure Smart Grid Network (BCS-SGN) has been created, utilizing group signatures and covert channel permission to guarantee user validity. Initially, the data from the smart grid network will be collected and encrypted using blowfish techniques. After encryption, the encrypted data will be stored in the block chain network, which also stores the transmission logs. In order to utilize the security controller, smart contracts are analyzed using smart grid devices. DES, AES, and BCS-SGN are widely used. Symmetric key algorithms are properly compared in this study. Additionally, the metrics of encryption/decryption time, computational time, and throughput are also compared. The percentage of the proposed method, BCS-SGN, is 20%, AES is 15%, and DES is 12%. These outcomes demonstrate that the proposed BCS-SGN outperforms other techniques.

Keywords *– Smart grid, Smart contract, Blowfish Algorithm, Block chain.*

1. INTRODUCTION

The recent blooming expansion of the network, in both wired and wireless environments, has been the primary driver of the smart grid as an emerging technology. As a result of the interconnection-based environment, there are many different types of edge infrastructure or devices available for energy service offers [1–3]. All electrical devices, smart meters, and other embedded systems with an energy-related focus are connected on the Smart Grid Network (SGN) platform [4]. A basic SGN can be set to have network nodes that are multiple electric sources and different user kinds [5–7]. As a result, network characteristics and service model usage might result in governance and optimization.

In order to guarantee sufficient confidentiality, integrity, and availability, researchers are encouraged to look into viable methods while using such a large-scale network. Users' privacy is a significant issue in addition to the usual security issues [8]. Data about power usage may be used to infer private information about users by entities with access to the data. Detecting household gadgets and people [9], profiling electric vehicles (EVs), and human location and activity patterns [10] are a few examples of this information.

Differential privacy is a notable and commonly used concept for formalizing data privacy [11]. It ensures that the presence or absence of an individual has no effect on the result by adding controlled randomness to the data. Moreover, "local" differential privacy is widely employed since it doesn't need a trustworthy data curator. Since blockchain doesn't save information on the identities of ENs or group signatures, the model protects privacy from the standpoint of the signature.

These methods have significantly advanced the protection of the privacy of power usage data [12]. Other techniques, such as homomorphic encryption and blockchain-based systems, have been utilized to protect the anonymity of energy usage [13]. This work makes an effort to offer a comprehensive system that secures and maintains user privacy in smart grids without relying on a centralized or reliable authority. The following list of contributions might be used to summarize this work.

- Initially the data from the smart grid network will be collected and encrypted using blowfish technique.
- After Encryption the encrypted data will be stored in the block chain network, which also stores the transmission logs. Smart contracts provide the most secure approach because they operate on the blockchain.
- This paper focused on privacy concerns related to smart grids and offered an alternative that would

enhance data security while maintaining smart grid performance.

- Additionally, the metrics of encryption/decryption time, computational time, and throughput are compared in order to evaluate the proposed method.

The remaining sections of the paper are arranged as follows. Section 2 offers a review of related work. The model design and the main suggested algorithms are then presented in Sections 3 and 4, respectively. Additionally, Section 5 provides analysis and evaluation outcomes. The profession is concluded in Section 6 lastly.

2. LITERATURE SURVEY

The literature on privacy-preserving smart grid systems has been reviewed. To deal with this problem, scholars have employed a variety of methods and resources.

In 2022 S., Jha, et.al [14] proposed a secure technique, in which the availability of energy, financial and environmental security has been interconnected that also affects the development of people. The electric power industry should place a strong emphasis on comprehensive energy security, which is based on the security of power grids. As a result, the conversion of energy for both essential and end uses is intricately interwoven, necessitating a large-scale energy plan.

In 2022 Subhash, P., et.al [15] proposed a big data framework, in the area of energy usage which has been influenced by the smart grid. The end-to-end, two-way delivery method offered by Smart Grid is effective and dependable. Real-time data on a user's individual electricity usage is provided. In order to get the star rating for each particular appliance, the study first determine the potential number of appliances used by a single user.

In 2021 Xiao, L., et.al [16] suggested a simple identity authentication technique, which uses elliptic curve encryption (ECC) technology, suitable for smart grid environments. In the suggested protocol, the identity and key information are encrypted using ECC, and the session's validity is checked using the timestamp. The cost analysis concludes by demonstrating that the suggested protocol is appropriate for implementation in large-scale intelligent smart grid setups since it has lower communication and calculation costs than other relevant protocols.

In 2021 Zainab et al [17] suggested big data management, in which order to handle the data in the grid. In order to comprehend the sources and types of data in the grid, data management tools and procedures have been used. The report highlights the shortcomings of the current approaches geared toward using large amounts of data from the smart grid.

3. BLOCK CHAIN-BASED SECURE SMART GRID NETWORK

In this study a novel Block Chain based Secure Smart Grid Network (BCS-SGN) has been proposed. Initially the data from the smart grid network will be collected and encrypted using blowfish techniques. After Encryption the encrypted data will be stored in the block chain network, which also stores the transmission logs. The proposed BCS-SGN protects and upholds user privacy without depending on a centralized authority. The overall planned workflow is shown in Figure 1.

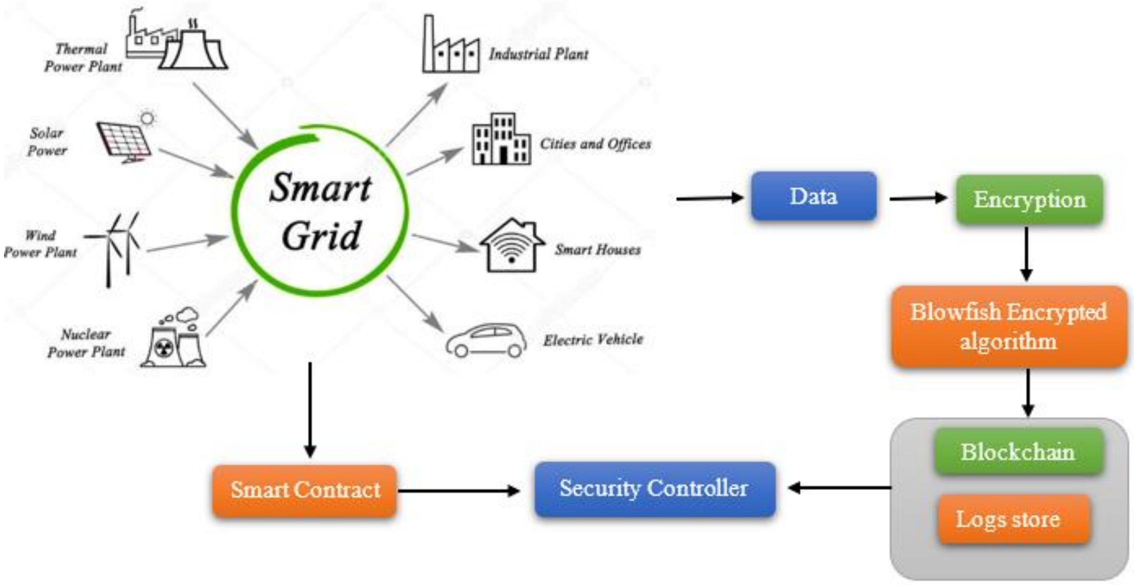


Figure 1. BCS-SGN framework

3.1 Blowfish Encrypted Algorithm

Using the same secret key for both encryption and decryption, Blowfish is a symmetric encryption method. Another block cipher is Blowfish, which separates a message into blocks of a predetermined length for encryption and decoding. Data encryption and protection are made possible with the help of the symmetric block cipher known as the Blowfish.

3.1.1 Data Encrypted Algorithm

The procedure repeats fifteen times, exchanging the two 32-bit segments (L and R) as the XOR operation is completed. The ciphertext is generated by performing an XOR calculation with the remaining P array and computing the cipher function following the fifteenth iteration. This algorithm, which uses S-boxes, is thought to have its most complicated component in the computation of the cipher function. The following algorithms are shown Table 1.

Table 1. Data Encrypted Algorithm

Algorithm 1: Data Encryption
Divide 64 bits in to two 32bit
Halves:1to 16
For i=1 to 16
XL=XL XOR Pi
XR=F(XL)XOR XR
Swap XL and XR
Swap XL and XR
XR=XR XOR P17
XL=XL XOR P18
Concatenate XL and XR

3.2 Block chain Smart Contract

The blockchain server where smart contracts are deployed is called the smart contract. Our model takes into account the practicality of applying blockchain technology so that a smart contract will determine the best course of action for energy allocations. We create dynamic programming in order to generate the best possible solution for the distribution of energy resources while taking into account three factors: communication security, latency time, and energy consumption. The smart contract is used to broadcast the operations, allowing each edge node to inform the node operator about how well it predicts it will handle the next request.

3.3 Block chain

A blockchain system that uses a layer of authorization to establish the scope of users or voters and grant access to the system to that target group is known as a permissioned blockchain. The platform for data storage that our model chooses is a permissioned blockchain for two main reasons. A key factor in the development of our approach is the potential of blockchain to protect privacy. According to our observations, a permissionless blockchain is not ideal considering that the user groups in the smart grid energy trading scenario are mostly internal entities connected via SGNs with a very constant identification state.

4. RESULT AND DISCUSSION

The performance of the Novel BlockChain-based Secure Smart Grid Network (BCS-SGN) has been discussed in this section. Matlab is one program that can be used to simulate the blockchain process. This can be used to spread the blockchain and mine blocks with incorrect hashes for testing, as multiple nodes can carry out the activity in the simulation. With Matlab, one gigabyte of RAM at minimum and sixty gigabytes of disk space at most are available for each worker. mimic procedures and perform MatLab algorithms on historical blockchain data. one gigabyte of RAM at minimum and sixty gigabytes of disk space at most for each worker. The assessment of the suggested methods performance is provided in this part. DES is a well-known algorithm, and the performance of AES was compared to the previously recommended methods, BCS-SGN.

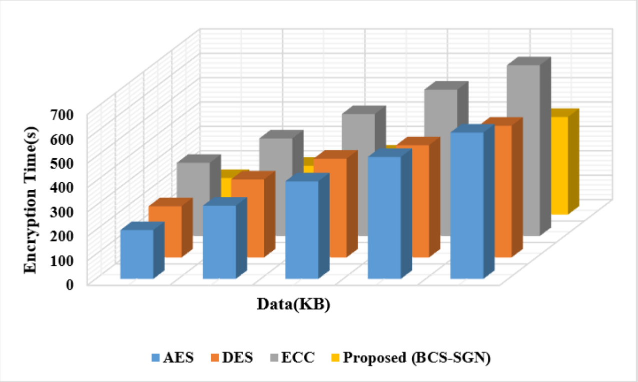


Figure 2. Encryption time

A graph comparing the encryption times the encryption part's graphical comparison is shown in Figure. 2. A nearly proposed approach from the Blowfish algorithm.

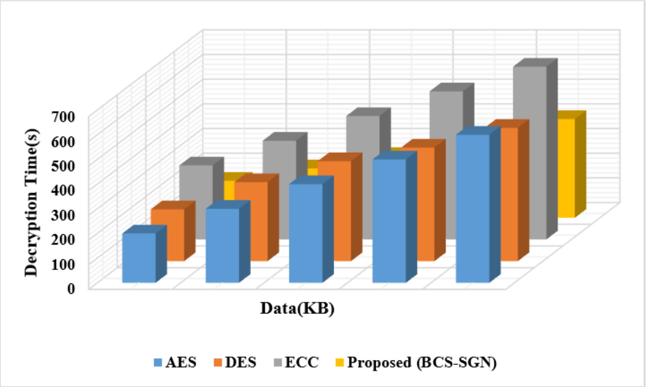


Figure 3. Decryption time

Figure 3 displays a graphical comparison of the decryption section. The suggested method outperforms the Blowfish algorithm.

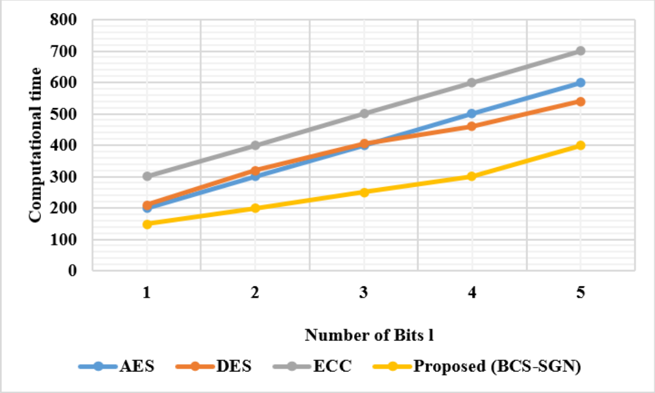


Figure 4. Computational time comparison

In Figure. 4, the bit number of the relevant characteristic, l, is directly correlated with the computational overheads of Extension.

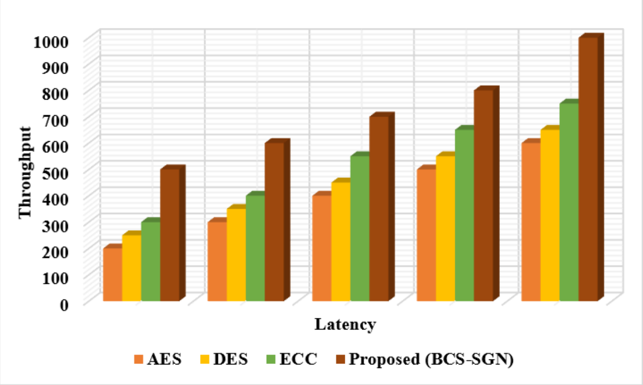


Figure 5. Throughput

In Figure. 5, the quantity of data or information moving from one location to another is known as throughput. By combining the total number of plaintexts in megabytes by the entire amount of time needed to complete each encryption step, the throughput of an encryption system can be computed.

5. CONCLUSION

In this paper, a novel blockchain-based smart grid network (BCS-SGN) has been created, utilizing group signatures and covert channel permission to guarantee user validity. Initially, the data from the smart grid network will be collected and encrypted using blowfish techniques. After encryption, the encrypted data will be stored in the block chain network, which also stores the transmission logs. In order to utilize the security controller, smart contracts are analyzed using smart grid devices. DES, AES, BCS-SGN, and the other four widely used symmetric key algorithms are properly compared in this study. Additionally, the metrics of encryption/decryption time, computational time, and throughput are also compared. The percentage of the proposed method, BCS-SGN, is 20%, AES is 15%, and DES is 12%. These outcomes demonstrate that the proposed BCS-SGN outperforms other techniques. This study concludes with a number of recommendations for additional research. The suggested algorithm needs to be improved upon or supported by additional research.

CONFLICTS OF INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

FUNDING STATEMENT

Not applicable.

ACKNOWLEDGEMENTS

The author would like to express his heartfelt gratitude to the supervisor for his guidance and unwavering support during this research for his guidance and support.

REFERENCES

[1] Y. Dai, D. Xu, S. Maharjan, Z. Chen, Q. He, and Y. Zhang, “Blockchain and deep reinforcement learning empowered intelligent 5g beyond”, *IEEE Network*, pp. 99, no. 1, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[2] S. Wang, X. Zhang, Y. Zhang, L. Wang, J. Yang, and W. Wang, “A survey on mobile edge networks: Convergence of computing, caching and communications”, *IEEE Access*, vol. 5, pp. 6757–6779, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[3] S. Maharjan, Q. Zhu, Y. Zhang, S. Gjessing, and T. Basar, “Dependable demand response management in the smart grid: A stackelberg game approach,” *IEEE Transactions on Smart Grid*, vol. 4, no. 1, pp. 120–132, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[4] N. Nikmehr and S. Ravadanegh, “Optimal power dispatch of multimicrogrids at future smart distribution grids”, *IEEE Transactions on Smart Grid*, vol. 6, no. 4, pp. 1648–1657, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[5] Y. Zhang, R. Yu, M. Nekovee, Y. Liu, S. Xie, and S. Gjessing, “Cognitive machine-to-machine communications: visions and potentials for the smart grid”, *IEEE Network*, vol. 26, no. 3, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[6] K. Wang, Y. Wang, X. Hu, Y. Sun, D. Deng, A. Vinel, and Y. Zhang, “Wireless big data computing in smart grid”, *IEEE Wireless Communications*, vol. 24, no. 2, pp. 58–64, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[7] Y. Zhang, R. Yu, S. Xie, W. Yao, Y. Xiao, and M. Guizani, “Home M2M networks: architectures, standards, and QoS improvement”, *IEEE Communications Magazine*, vol. 49, no. 4, pp. 44–52, 2011. [CrossRef] [Google Scholar] [Publisher Link]

[8] Y. Cai, Y. Cao, Y. Li, T. Huang, and B. Zhou, “Cascading failure analysis considering interaction between power grids and communication networks”, *IEEE Transactions on Smart Grid*, vol. 7, no. 1, pp. 530– 538, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[9] K. Gai, M. Qiu, Z. Ming, H. Zhao, and L. Qiu. Spoofing-jamming attack strategy using optimal power distributions in wireless smart grid networks. IEEE Transactions on Smart Grid, vol. 8, no. 5, pp. 2431– 2439, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[10] G. Liang, J. Zhao, F. Luo, S. Weller, and Z. Dong, “A review of false data injection attacks against modern power systems”, *IEEE Transactions on Smart Grid*, vol. 8, no. 4, pp. 1630–1638, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[11] K. Gai, M. Qiu, Z. Ming, H. Zhao, and L. Qiu, “Spoofing-jamming attack strategy using optimal power distributions in wireless smart grid networks”, *IEEE Transactions on Smart Grid*, vol. 8, no. 5, pp. 2431– 2439, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[12] P. McDaniel and S. McLaughlin, “Security and privacy challenges in the smart grid”, *IEEE Security & Privacy*, vol. 3, no. 75–77, 2009. [CrossRef] [Google Scholar] [Publisher Link]

[13] M. Khan and K. Salah, “Iot security: Review, blockchain solutions, and open challenges”, *Future Generation Computer Systems*, vol. 82, pp. 395– 411, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[14] S. Jha, H.A. Abdeljaber, M.K. Imam Rahmani, M.M. Waris, A. Singh, and M. Yaseen, “An Integration of IoT, IoC, and IoE towards building a green society”, *Scientific Programming*, 2022 [CrossRef] [Google Scholar] [Publisher Link]

[15] P. Subhash, K.S. Surya, and A.B. Reddy, “Analysis of Smart Grid Data for Appliance Prediction and Efficient Power Consumption”, *In 2022 International Conference on Recent Trends in Microelectronics, Automation, Computing and*

Communications Systems (ICMACC) pp. 1-5, 2022. IEEE. [CrossRef] [Google Scholar] [Publisher Link]

[16] L. Xiao, J. Cai, M. Qiu, and M. Liu, "A Secure Identity Authentication Protocol for Edge Data in Smart Grid Environment", *In 2021 8th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/2021 7th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom),* pp. 188-193, 2021. IEEE. [CrossRef] [Google Scholar] [Publisher Link]

[17] A. Zainab, A. Ghrayeb, D. Syed, H. Abu-Rub, S.S. Refaat, and O. Bouhali, "Big data management in smart grids: Technologies and challenges", *IEEE Access,* vol. 9, pp.73046-73059, 2021. [CrossRef] [Google Scholar] [Publisher Link]

AUTHORS

R.C. Ilambirai completed her B.E degree in the Department of Electrical and Electronics Engineering in 2002, M.Tech in Power Electronics in 2004 and her research in 2022 all from Tamilnadu, India.She is working as an Assistant Professor in the Department of Electrical and Electronics Engineering, SRM Institute of Science and Technology, Chennai, Tamilnadu, India.She has a good expertise in the field of inverters, dc-dc power converters, integration of converters with renewable energy sources, their applications etc. Overall, she has published 18 papers in International Journals and 8 Conference publications. She is a life member of MISTE, MIEI and member of IEEE.

S. Lourdu Jame is working as an Assistant Professor in the Department of Electrical and Electronics Engineering in SRM Institute of Science and Technology. She completed her B.E Electrical and Electronics Engineering in the year 2010, M.Tech in the year 2012 and Ph.D in the year 2023. The total professional experience is 11 years. Her areas of specialization are Thermoelectric Generator, Energy storage, Energy conservation. She has published 12 international publications among which 2 are SCI indexed journals. She is a life member of IEEE.

P.U. Poornima obtained her B.Tech in the Electrical and Electronics Engineering in 2003 under JNTU, Hyderabad and M.Tech in Power Electronics in 2005 from VIT University, Vellore and Ph.d in 2022 from SRMIST, Tamilnadu, India. She is working as an Assistant Professor in Electrical &Electronics Engineering, SRM Institute of Science and Technology, Chennai. Her research areas of interests are Integration of renewable energy sources, Inverters, Converters, Drives applications etc. She has published 10 research papers in international journals and presented four papers in International and National conferences. She is a life member of IEI and member of IEEE.

Arrived: 05.05.2024
Accepted: 17.06.2024

International Journal of Data Science and Artificial Intelligence (IJDSAI)
Volume 02, Issue 03, May – June (2024)

RESEARCH ARTICLE

SEMANTIC FEATURE ENABLED AGGLOMERATIVE CLUSTERING FOR INFORMATION TECHNOLOGY JOB PROFILE ANALYSIS

B. Jaison [1, *], R. Gladis Kiruba [2] and G Belshia Jebamalar [3]

[1] Department of Computer Science and Engineering, R.M.K. Engineering College, Kavaraipettai-601206 India.
[2] Department of Electronics and Communication Engineering and AIML, Bangalore College of Engineering and Technology, Chandapura, Bengaluru, India.
[3] Department of Computer Science Engineering, S.A Engineering College, Thiruverkadu, Tamil Nadu 600077 India.

Corresponding e-mail: bjn.cse@rmkec.ac.in

Abstract – The maintenance and implementation of computer systems are the core activities of information technology. Database administration and network architecture are also included in information technology. Professionals have access to a working environment that facilitates the setup of internal networks and the development of computer systems. There is an immediate need for a suitable approach to close the gap between supply and demand for IT workers. Extensive research into IT job profiles is crucial to meeting industry demands. Educational programs must identify the abilities that the industry requires to modernize its manufacturing. Semantic Feature-Enabled Agglomerative Clustering for Information Technology Job Profiling (SEA-IT) has been proposed to overcome these challenges. Semantic analysis is performed using a tree-like strategy. The most frequently used phrases and words from each cluster of IT professions were collected to demonstrate specific knowledge. Initially, the data from the online job posting sources will be collected and pre-processed using techniques such as stemming, normalization, text correction, removing stop words, and tokenization. Secondly, the pre-processed data can extract features using a bag of words. After feature extraction, the cluster is generated using an agglomerative algorithm to form an IT job analysis result, so that the knowledge and capabilities of IT professionals can be upgraded. The simulation findings, based on evaluation criteria and other statistical tests, demonstrated the suggested algorithm. Experiments demonstrated that SEA-IT functions well with a variety of descriptive methodologies and is independent of the dataset's dimensions.

Keywords *– Information Technology, Preprocessing, bag of words, agglomerative algorithm.*

1. INTRODUCTION

Informatics engineering is a subject that teaches students the principles of computer science and mathematical analysis in order to create, build, test, and evaluate software [1]. Database administrators, data scientists, UI designers, IT project managers, network engineers, UX designers, and other professions in the field of informatics technology are all open to graduates of the Department of Informatics Engineering with a computer science degree [2]. It differs from other majors in that a student with a doctoral degree in education will work as a doctor, a student with a nursing degree will work as a nurse, a student with a pharmacy degree will work as a pharmacist, a student with an education degree will work as a teacher, and so on [3].

However, students studying informatics engineering won't be sufficient if they merely have a good education without complementing talents [4]. Students majoring in informatics engineering must broaden their skill sets in order to meet the requirements for expert workers in certain IT domains [5]. Career seekers must be aware of the specialized knowledge and skills required for each IT career field from the wide variety of IT job fields [6]. The scarcity of graduates prepared for the workplace in informatics technology areas is caused by the fact that informatics engineering students frequently do not know the benchmark of their capacity for work skill requirements in IT industries [7].

Application system suggestion careers in IT sectors based on IT knowledge and abilities are therefore essential [8]. Even though each student who completes the Informatics Engineering program is equal, they all have unique knowledge and talents [9]. The terms and phrases that are most frequently used in the information technology sector to represent skills and knowledge as a new dataset serve as the foundation for the model's functionality. Ten IT professionals from various commercial and government enterprises in Indonesia confirmed the skills needed for each position through focus group discussions (FGD) [10]. The main contribution of the suggested method is as follows:

- Initially, the data from the online job posting sources will be collected and pre-processed using techniques such as stemming, normalization, text correction, removing stop words, and tokenization.

- Secondly, the pre-processed data can extract features using a bag of words. After feature extraction, the cluster is generated using an agglomerative algorithm to form an IT job analysis result. so that the knowledge and capabilities of IT professionals can be upgraded;
- The simulation findings, based on evaluation criteria and other statistical tests, demonstrated the suggested algorithm.
- Experiments demonstrated that SEA-IT functions well with a variety of descriptive methodologies and is independent of the dataset's dimensions.

The remaining sections of the paper are arranged as follows. Section 2 offers a review of related work. The model design and the main suggested algorithms are then presented in Sections 3 and 4, respectively. Additionally, Section 5 provides analysis and evaluation outcomes. The profession is concluded in Section 6 lastly.

2. LITERATURE SURVEY

In 2022 E. Novak et al. [11] proposed to identify the job profiles that IT specialists need. A systematic semantic technique was suggested using a hierarchical clustering analysis based on average linkage. Semantic analysis, which is akin to a tree structure technique, is used to uncover pertinent phrases, connections, and hidden meanings. The end result is a methodical semantic examination of the programming language, specialized kind, task, database, tools, and frameworks included in the IT job profile. Ten IT experts from various government and commercial enterprises in Indonesia participated in focus groups (FGD) to confirm the rationale behind each job profile.

In 2020 T. Bai, et al. [12] proposed the users should be able to retrieve needed information with ease thanks to the depiction and structuring of information elements. Thus, through a thorough literature review and the administration of two surveys, one for IT employers and the other for graduates, this paper seeks to determine the key variables influencing IT graduates' employability and capacity to compete in local, regional, and global labor markets. Then, using the Statistical Package for Social Sciences (SPSS) 28.0, data were gathered and examined in order to construct our suggested framework. This framework would incorporate all relevant variables and stakeholders in order to improve graduates' employability and align with market expectations.

In 2021 K. Binici, et al. [13] suggested to determine the information technology competencies needed in information institutions. Information science and library technology developers frequently use the code4lib platform, which is where the study's data was gathered. Among the outcomes is a list of information technology competencies required in information institutions, especially for technologists, instructors, and aspiring information workers.

In 2022 Mehirig, A., et al. [14] suggested a brand-new, analytical approach that is totally reproducible, semi-automated, and built on a blend of expert judgment and machine learning algorithms. In this approach to creates a comprehensible classification of job responsibilities and skill sets by utilizing a sizable volume of online job ads that were gathered through web scraping. The findings can help HR managers and corporate executives create clear plans for acquiring and developing the skills necessary to fully utilize big data.

3. PROPOSED METHODOLOGY

3.1 Information Technology jobs

In this study a Novel Semantic Feature Enabled Agglomerative Clustering for Information Technology job profiling (SEA-IT) has been proposed.

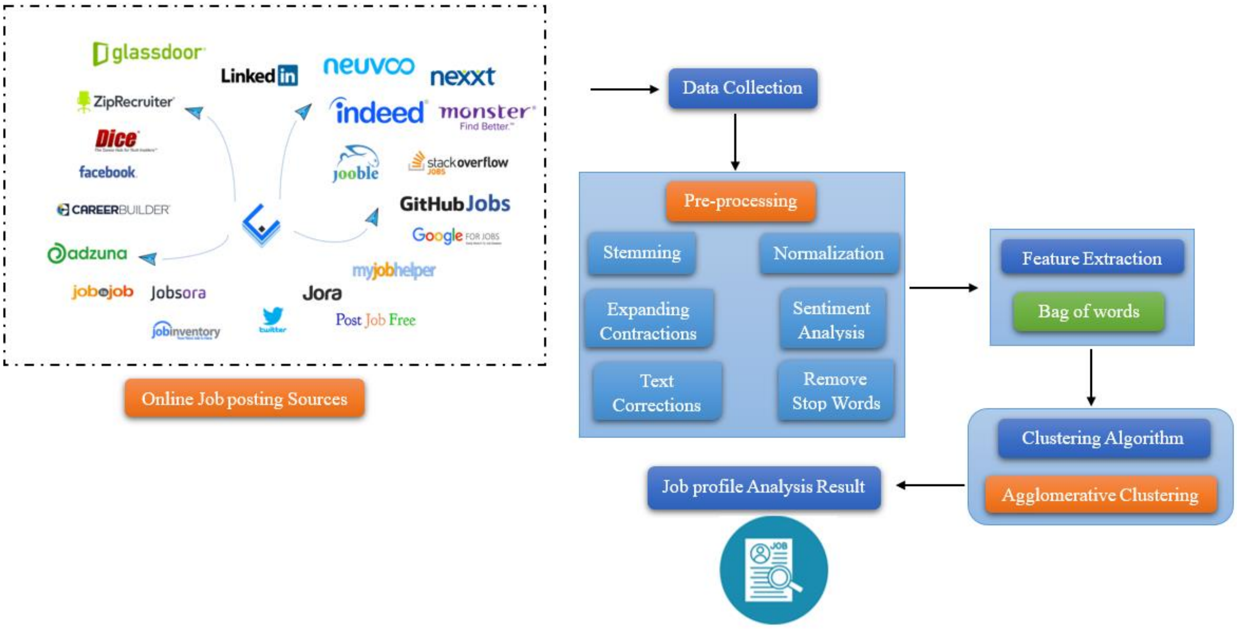


Figure 1. Block Diagram of Proposed methodology

Initially, the data from the online job posting sources will be collected and pre-processed using techniques such as stemming, normalization, text correction, removing stop words, and tokenization. Secondly, the pre-processed data can be utilized to extract features using a bag of words. After feature extraction, the cluster is generated using an agglomerative algorithm to form an IT job analysis result, so that the knowledge and capabilities of IT professionals can be upgraded. The overall planned workflow is shown in Figure 1.

3.2 Data Collection

Web-based actual datasets have gained popularity lately in information retrieval studies. IT job postings are the most significant and engaging source of new hires. They have typically been distributed through newspapers, but in the current digital era, they are primarily published online, on specialist websites or social media. Numerous research has made use of job adverts from sites including Bebee, Naukri.com, CareerBuilder.com, and LinkedIn. For a number of reasons, this study investigated the Job Street Indonesia and Tech in Asia platforms. They start by concentrating on IT job postings in Asia. Secondly, they provided a talent tag for each job posting that was advertised. Lastly, Tech in Asia supports the tech and entrepreneurial communities in Asia and has a catchy tagline.

3.3 Preprocessing

To guarantee accurate data processing, pre-processing must to be carried out on unstructured databases. This study involved six pre-processing techniques, as shown in below.

- **Stemming:** This process, often referred to as text standardization, involves stemming or reducing words to their most basic or root form.
- **Normalization:** Textual data may contain several words with comparable meanings. Normalization is the act of reducing the various forms of words with comparable meanings to a single standard form in order to standardize literature. By reducing the amount of randomness, this procedure seeks to enhance text quality, lower vocabulary size, boost text processing effectiveness, and optimize NLP model performance.
- **Text Correction:** Text correction is the process of fixing spelling and grammatical problems in order to ensure consistency with terms in a text and avoid different vector representations for misspelled words.
- **Remove Stop Words:** Words that add nothing to the meaning of a sentence are known as stop words. Consequently, their removal would not alter the sentence's meaning.
- **Tokenization:** Tokenization, in natural language processing, is the act of splitting texts into smaller word or sentence portions known as tokens. Tokenization divides text into discrete pieces so that models can identify patterns within them. This helps the model comprehend and represent the content more accurately.

3.4 Feature Extraction Using Bag of Words Algorithm

Dense SIFT features that are obtained from every SDP image using Euclidean distance are used for visual word matching. A numerical vector with a K-dimensional representation of the fault sample is created by counting the frequency of a specific type of word in the image. A hierarchical pyramid structure is formed by breaking down all different types of frequency vectors in picture words into blocks using spatial pyramid matching. This structure may be thought of as a bag of words with spatial-scale features.

3.5 Agglomerative Clustering Algorithm

Among the most well-liked and successful techniques for grouping data are agglomerative methods. However, a comprehensive comparison of these approaches concerning the crucial problem of inaccurate results during cluster search has not been conducted. A cluster model with a higher density centre encircled by a transition and outliers is used to measure the significance of the discovered clusters and deal with the false positive issue.

```
Algorithm 1: Pseudocode of the agglomerative hierarchical clustering

Input: Dataset M with S instances as $M = \{X_1, X_2 \dots\dots X_S\}$ and cluster distance function L$(d_i, d_j)$

Output: Black partition dendrogram Z each 1$\leq Z \leq S$

$d_i$ = {$X_i$},$\forall_i$=1, 2…S
For Z=S down to 1 do
    $Dendrogram_Z$={$d_1, d_2, \dots d_S$}
    L (i, j) = L ($L_i, L_j$) ,$\forall_i$= 1,2, …. S
    Let (z, f) = $argmin_{(i,j)}\{L(d_i, d_j): 1 \leq i \leq j \leq Z + 1\}$
    CZ=Join ($C_Z, C_f$)
    Remove ($C_f$)
End
```

3.6 Job profile Analysis

The median-LinkedIn certain visualizations, such as word/phrase frequency analysis and link analysis, which describe words and phrases connected together, use the hierarchical clustering technique. This study's employment profile analysis was explained using these visualization concepts. The most frequently used terms in the textual corpus can be quickly viewed through word or phrase frequency analysis. Finding keywords that belong to the same clusters and the related terms that may be utilized to get information about these clusters is also beneficial. This function might also aid in the identification of co-occurring words, enabling the investigation of search phrases to find pertinent text. A word cloud graphical display and table format can be used to present a visual result. Each word's different sizes are displayed proportionately to how frequently it appears in the text. A link analysis, which shows the relationship between keywords, is displayed as a network graph. Based on the experiment conducted, this study has several benefits. For instance, the number of clusters that can be specified can be changed based on the job profile that needs to be investigated; it can run quickly computationally; the graphical representation makes it easier for readers to understand; and different cluster sizes and shapes can be chosen to meet the objectives of the study.

4. RESULT AND DISCUSSION

The performance of the Novel Semantic Feature Enabled Agglomerative Clustering for Information Technology Job Profiling (SEA-IT) has been discussed in this section. Matlab is one program that can be used to simulate the blockchain process. This can be used to spread the blockchain and mine blocks with incorrect hashes for testing, as multiple nodes can carry out the activity in the simulation. With Matlab, one gigabyte of RAM at the minimum and sixty gigabytes of disk space at most are available for each worker. mimic procedures and perform Matlab algorithms on historical blockchain data. one gigabyte of RAM at a minimum and sixty gigabytes of disk space at most for each worker.

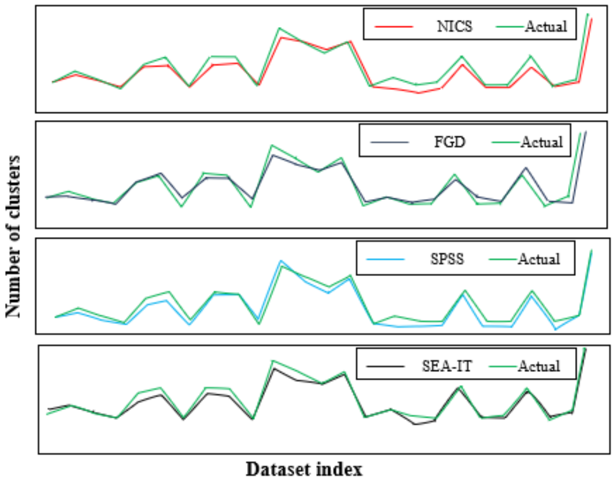


Figure 2. Comparison of SEA-IT with Existing Algorithms in the number of optimal clusters

Figure. 2 displays the comparison findings between the ensemble clustering and SEA-IT algorithms. For every dataset, two counts are provided: the number of clusters found by the clustering algorithms and the number of clusters in reality. The results unequivocally demonstrate that the SEA-IT curve and the curve corresponding to the number of real clusters diverge less, with the SPSS algorithm producing superior results. In SEA-IT, the average discrepancy between the number of clusters and the actual number of clusters is 15.5. The variations are documented for the algorithms, which are 16.9, 15.4, and 15.0, in that order.

Figure 3 presents the comparison findings for several algorithms. The number of dataset instances that are accurately allocated to the appropriate class determines the clustering accuracy. As a result, the ratio of correctly grouped instances to total instances is known as clustering accuracy. As demonstrated, SEA-IT provides more accuracy than other algorithms in most regards. In terms of data clustering, SEA-IT outperforms, SPSS, FGD, and NICS by an average of 5.2%, 2.3%, and 1.0%, respectively.

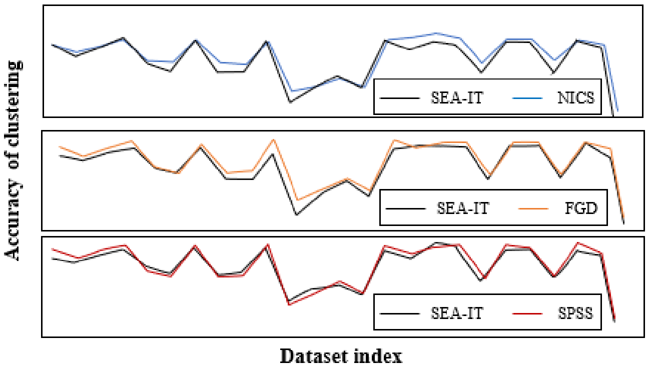


Figure 3. Comparison of SEA-IT with existing Algorithms in clustering Accuracy

5. CONCLUSION

In this research, a novel Semantic Feature Enabled Agglomerative Clustering for Information Technology job profiling (SEA-IT) has been proposed. Semantic analysis is performed using a tree-like strategy. The most frequently used phrases and words from each cluster of IT professions were collected to demonstrate specific knowledge. Initially, the data from the online job posting sources will be collected and pre-processed using techniques such as stemming, normalization, text correction, removing stop words, and tokenization. Secondly, the pre-processed data can extract features using a bag of words. After feature extraction, the cluster is generated using an agglomerative algorithm to form an IT job analysis result, so that the knowledge and capabilities of IT professionals can be upgraded. The simulation findings, based on evaluation criteria and other statistical tests, demonstrated the suggested algorithm. Experiments demonstrated the SEA-IT functions well with a variety of descriptive methodologies and is independent of the dataset's dimensions.

CONFLICTS OF INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

FUNDING STATEMENT

Not applicable.

ACKNOWLEDGEMENTS

The author would like to express his heartfelt gratitude to the supervisor for his guidance and unwavering support during this research for his guidance and support.

REFERENCES

[1] O. Semenikhina, V. Proshkin, and O. Naboka, “Application of Computer Mathematical Tools in University Training of Computer Science and Mathematics Pre-service Teachers”, *International Journal of Research in E-learning*, vol. 6, no. 2, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[2] L.D. Kumalasari, and A. Susanto, “Recommendation system of information technology jobs using collaborative filtering method based on LinkedIn skills endorsement”, *Sisforma*, vol.

6, no. 2, pp.63-72, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[3] T.S. Prøitz, and L. Wittek, "New directions in doctoral programmes: bridging tensions between theory and practice?", *Teaching in Higher Education*, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[4] J. Qadir, and A. Al-Fuqaha, "A Student Primer on How to Thrive in Engineering Education during and beyond COVID-19", *Education Sciences,* vol. 10, no. 9, pp. 236, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[5] L. Vyas, and N. Butakhieo, "The impact of working from home during COVID-19 on work and life domains: an exploratory study on Hong Kong", *Policy design and practice*, vol. 4, no. 1, pp. 59-76, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[6] S. More, and T. Rosenbloom, "Job-field underemployment and its impact on the demand for higher education at the Israeli labor market", *Israel Affairs*, vol. 28, no. 2, pp. 316-334, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[7] X. Xu, Y. Lu, B. Vogel-Heuser, and L. Wang, "Industry 4.0 and Industry 5.0—Inception, conception and perception", *Journal of manufacturing systems*, vol. 61, pp. 530-535, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[8] M.E. Armstrong, K.S. Jones, A.S. Namin, and D.C. Newton, "Knowledge, skills, and abilities for specialized curricula in cyber defense: Results from interviews with cyber professionals", *ACM Transactions on Computing Education (TOCE)*, vol. 20, no. 4, pp.1-25, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[9] M. Torres, N. Flores, and R. Torres, "Fostering soft and hard skills for innovation among informatics engineering students: An emancipatory approach", *Journal of Innovation Management,* vol. 8, no. 1, pp. 20-38, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[10] J. Hammerschmidt, S. Durst, S. Kraus, and K. Puumalainen, "Professional football clubs and empirical evidence from the COVID-19 crisis: Time for sport entrepreneurship?", *Technological Forecasting and Social Change*, vol. 165, pp. 120572, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[11] E. Novak, L. Bizjak, D. Mladenić, and M. Grobelnik, "Why is a document relevant? Understanding the relevance scores in cross-lingual document retrieval", *Knowledge-Based Syst.*, vol. 244, pp. 108545. [CrossRef] [Google Scholar] [Publisher Link]

[12] T. Bai, Y. Ge, S. Guo, Z. Zhang, and L. Gong, "Enhanced Natural Language Interface for Web-Based Information Retrieval," *IEEE Access*, vol. IX, pp. 4233–4241, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[13] K. Binici, "What are the information technology skills needed in information institutions? The case of 'code4lib' job listings", *J. Acad. Librariansh.*, vol. 47, no. 3, pp. 102360, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[14] A. Mehirig, "Skills required in Big Data professions", *Journal of Advanced Economic Research//V,* vol. 7, no. 01, 2022. [CrossRef] [Google Scholar] [Publisher Link]

AUTHORS

B. Jaison is currently working as Professor in the Department of Computer Science and Engineering, RMK Engineering College, Kavarapettai, Chennai, India. He is having more than 24 years of teaching Experience in different Institutions in various cadres. He completed his M.E degree in Computer Science & Engineering from G.K.M College of Engineering and Technology, Affiliated to Anna University, Chennai in the year 2007 and Ph.D in Information and Communication Engineering from Anna University, Chennai in the Year 2015. He has published more than 50 Research Articles in International Journals and attended many International Conferences. He is the recognized Research Supervisor of Anna University Chennai and produced Five Research Scholars. His areas of interest include Data mining, Image Processing and Cloud Computing. He is a life member in IAENG, IACSIT and ISTE.

R. Gladis Kiruba received her UG degree in B. E (ECE) in PSN college of Engineering and Technology during the year 2004-2008 and received her master degree MTech (VLSI design) in karunya university during the year 2008-2010.She is currently working as Assistant professor in Bangalore college of Engineering and technology, chandapura, Bangaluru in the year April (2024).

G Belshia Jebamalar is currently working as Assistant professor in SA Engineering college in Department of computer science in the year 2024.

Arrived: 18.05.2024

Accepted: 25.06.2024