Enhancing Privacy-Preserving Intrusion Detection in Blockchain-Based Networks with Deep Learning

RESEARCH PAPER

JUNZHOU LI
QIANHUI SUN
FEIXIAN SUN

*Author affiliations can be found in the back matter of this article

ABSTRACT

Data transfer in sensitive industries such as healthcare presents significant challenges due to privacy issues, which makes it difficult to collaborate and use machine learning effectively. These issues are explored in this study by looking at how hybrid learning approaches can be used to move models between users and consumers as well as within organizations. Blockchain technology is used, compensating participants with tokens, to provide privacy-preserving data collection and safe model transfer. The proposed approach combines Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) to create a privacy-preserving secure framework for predictive analytics. LSTM-GRU-based federated learning techniques are used for local model training. The approach uses blockchain to securely transmit data to a distributed, decentralised cloud server, guaranteeing data confidentiality and privacy using a variety of storage techniques. This architecture addresses privacy issues and encourages seamless cooperation by utilising hybrid learning, federated learning, and blockchain technology. The study contributes to bridging the gap between secure data transfer and effective deep learning, specifically within sensitive domains. Experimental results demonstrate an impressive accuracy rate of 99.01%.

CORRESPONDING AUTHOR:
Feixian Sun
School of Electronics and Internet of Things of Henan Polytechnic, Zhengzhou 450046, CN
feixiansun@163.com

KEYWORDS:
blockchain; LSTM; GRU; privacy

TO CITE THIS ARTICLE:
Li, J, Sun, Q and Sun, F. 2023. Enhancing Privacy-Preserving Intrusion Detection in Blockchain-Based Networks with Deep Learning. *Data Science Journal*, 22: 31, pp. 1–11. DOI: https://doi.org/10.5334/dsj-2023-031

1 INTRODUCTION

To build trust with customers who prioritize security, organizations must prioritize transparency in three key areas: gaining permission to store private data, following privacy regulations, and organizing the collected data. By being transparent in these aspects, organizations can demonstrate their commitment to protecting customer data and complying with legal requirements. Failure to adhere to these regulations can result in significant fines and reputational damage (McMahan & Ramage 2017). In addition to legal compliance, organizations need to address the potential risks posed by external threats like malware or hackers. These attacks can lead to financial losses and a loss of customer confidence. To mitigate such risks and maintain customer trust, organizations must adopt transparent practices in data collection, data handling, and data security measures (Benhar, Idri & Fernández-Alemán 2020). When it comes to data privacy in the context of machine learning, the Differential Privacy technique plays a crucial role. It ensures that individual identities within a dataset remain anonymous, preventing viewers from associating specific individuals with the results. By introducing random noise through a distribution, the technique protects individuals' privacy by obscuring their genuine answers (Dwork 2011; Li, Li & Varshney 2022). To ensure accurate and reliable data analysis, organizations can employ data cleaning techniques and leverage various machine learning frameworks that offer methods and APIs for data imputation. Missing values can be assigned using statistical measures such as medians, means, standard deviations, or utilizing techniques like k-nearest neighbors (k-NNs) (Chollet 2017; Choudhury et al. 2020). Machine learning algorithms, such as Support Vector Machines (SVM), clustering, and neural networks, are employed to analyze and find patterns in large datasets. Clustering allows the grouping of similar data pieces, enabling the discovery and examination of patterns across multiple datasets (Domadiya & Udai Pratap 2021; Hamza & Muhammad 2020). Neural networks, inspired by cognitive processes, excel at identifying complex patterns (He et al. 2016a; He et al. 2016b). Convolution is a technique used in neural networks where each layer focuses on specific features, gradually capturing higher-level properties (Howard et al. 2017). To perform secure computations on encrypted data, organizations can utilize homomorphic encryption. This technique enables calculations to be carried out on encrypted data, producing results that resemble what would have been obtained from plaintext data (Bos, Lauter & Naehrig 2014; Huang et al. 2017).

Federated learning addresses the challenge of handling heterogeneous data. Instead of locally storing or transmitting raw data, individual client's data remains private. Analysts aggregate client data instead of accessing specific communications, ensuring privacy while enabling rapid analysis (Weng et al. 2019).

While differential privacy protects individual privacy, it introduces a tradeoff with accuracy, such as differential privacy techniques introduce noise or perturbations to protect individual privacy, but this can impact the accuracy of the analysis and make it challenging to draw conclusions from individual samples. Alterations made during the randomization process slightly impact the outcome distribution, making it difficult to draw conclusions from individual samples. Gradient-based learning systems can achieve differential privacy by introducing random perturbations to intermediate outputs, such as using Gaussian noise (Zhao, Chen & Zhang 2019). In federated learning, FedAvg is a distributed averaging approach that ensures communication efficiency. It involves training multiple clients and aggregating their models to achieve the desired outcome (Jamil & Kim 2021). Federated learning distinguishes between local and global privacy. Global privacy ensures modifications made to the model at each round remain secret from all external parties, preserving worldwide anonymity. Local privacy safeguards changes from being visible to the server as well. Minimizing data on the server helps reduce memory and computation requirements during training iterations, which is known as data normalization (Jena & Debaprada 2021).

Blockchain technology provides a solution to reduce privacy erosion while enabling controlled data sharing. Users can selectively disclose parts of their personal data on a blockchain to access specific services. The transparency and decentralization of blockchain, exemplified by cryptocurrencies like Bitcoin, have demonstrated their reliability in managing information (Krizhevsky & Sutskever 2012; Yin et al. 2021).

Consensus algorithms play a crucial role in ensuring agreement among participants in a blockchain network. These algorithms enhance network stability and foster decentralized trust among anonymous peers (Li et al., 2020). Proof of work is a cryptographic mechanism that demonstrates a participant's computational effort, while proof of stake selects stakeholders based on their holdings of the cryptocurrency involved (Li et al., 2021).

The main objective of this work is to enhance privacy-preserving techniques in intrusion detection systems deployed in blockchain-based networks. Leveraging federated deep learning, the proposed model ensures efficient and accurate intrusion detection while maintaining data privacy. By utilizing federated learning techniques, individual client data remains confidential, allowing collaborative model training and analysis.

2 LITERATURE SURVEY

Qiang, Liu & Jin (2021) indicate that convolutional neural networks (CNNs) and binary neural networks (BNNs) can be used for collecting data, encrypting it before storing it in the cloud, and training and testing without additional decryption. Their system might be dangerous since they store all of their data in one place before encrypting it. Iterative search technique was used by Rui Hu et al. (2020) in order to propose a personalised federated learning strategy that offers strong privacy for user data regardless of user heterogeneity. However, because of the heterogeneity of the devices, this method's training process is very difficult and complex. Data science and machine learning methods such as homomorphic encryption and dimensionality reduction can be utilized to guarantee the confidentiality of data. Rahman et al. (2020) presented such a system by which dimensionality reduction and homomorphic encryption can be used to guarantee data confidentiality. The suggested method is made to provide users confidence that machine learning would be used to preserve their data privacy and prevent their personal information from being used for commercial gain. Only certain illnesses and medical conditions are treated using this technique. A machine learning strategy was described by S. Shaham et al. (2021) with the aim of releasing the location of data while maintaining anonymity. Its strategy incorporates K-means algorithms as well as clustering, alignment, and generalization methods. By using MLA, users' privacy is protected while geographical itinerary datasets can be made available, an anonymization framework based on machine learning. Tanwar et al. (2020) methodology for creating intelligent blockchain-based apps has been described. It involves the use of machine learning methods. Secure Hash Algorithm (SHA), the consensus algorithm, is used in this process. A smart city, healthcare system, smart grid, or unmanned aerial vehicle (UAV) could all benefit from this methodology, which combines machine learning and blockchain technology. Due to a high demand for internet bandwidth and an increase in chain, performance appears to be hindered. The privacy of datasets can be modified based on the distribution of data, according to Wang et al. (2019). Government transmission, storage, and learning training efficiency has improved as well as the security of client data. Sparse differential gradients increase gearbox efficiency, but their accuracy declines by 0.03%. Using heterogeneous data rather than homogeneous data, Decentralised Federated Learning through Mutual Knowledge Transfer is proposed by Shayan et al. After a certain number of cycles, this technique is more accurate than baseline techniques. For this approach to perform better, additional theoretical research has to be done. Based on a variety of datasets, experimental setups, and privacy budgets, Simonyan (2014) found that logistic regression had better performance than differential regression. It has been discovered, however, that differential privacy causes significantly worse performance degradation in federated learning. In a framework called PPSF (Srivastava et al. 2021), Srivastava et al. (2021) propose IoT-driven smart cities employing blockchain and machine learning. Based on LightGBM, e-PoW, and Principal Component Analysis (PCA), this system can be used to perform the analysis. With PPSF, smart cities powered by IoT can maintain privacy and security using blockchain technology and machine learning. Without the need of a centralized model coordinator, Szegedy et al. (2016) developed a decentralized, trustworthy, and secure technique for federated learning. This improves model update privacy security and successfully thwarts concerns of data poisoning. A slower convergence rate is observed for the proposed model compared with the SAE model. A strategy for implementing design using the present blockchain technology and compensating employees with bitcoin for following the protocols was put out by Toyoda, Zhao & Zhang (2020).

While this solution does not use blockchains in its implementation, it does use an open network. A technique for differential privacy publication of a medical data model was proposed by Z. Sun et al. (2019). Algorithms including Mini Batch Gradient Descent (MBGD), Differentially Private Mini Batch (DPMB), Gradient Descent (GD), and Back Propagation (BP) are used by their prosed system. When it comes to releasing and training data, it offers adequate privacy guarantees and privacy protection. This approach relies on a small number of datasets, so using several datasets at once reduces accuracy. Zhao et al. (2019) have developed a system that utilizes a stochastic gradient descent algorithm, a method for generating communication policies, and heterogeneous networks based on machine learning approaches to increase communication effectiveness through decentralized networks. NetMax is a decentralized and communication-efficient method for accelerating distributed machine learning over heterogeneous networks. Their system malfunctions when a primary server is present, which raises concerns about data privacy. As in the literature deep learning based approaches are employed to secure environement (Alatawi & Mohammed 2023; Saba et al. 2022), CNN is the most common deep learning approach (Yar et al. 2021). The existing literature primarily focuses on individual privacy-preserving techniques or strategies, neglecting the potential synergies that can be achieved by combining multiple approaches. However, it is crucial to explore the benefits and challenges associated with integrating various privacy-preserving methods to provide stronger assurances of data confidentiality and privacy. There is a research gap in understanding how these different techniques can work together cohesively to enhance overall privacy protection. Therefore, there is a need for further investigation to explore the potential advantages and complexities of leveraging a holistic approach that combines multiple privacy-preserving approaches. By addressing the research gap, the aim is to provide more robust and comprehensive solutions for ensuring data confidentiality and privacy in this study.

3 PROPOSED MODEL

The system model depicted in Figure 1 demonstrates the process of data sharing using a combination of federated learning and blockchain technology. The data requester initiates the process by publishing a task on the blockchain, indicating the need for data sharing. Relevant data nodes receive this request and respond accordingly. Through consensus mechanisms,

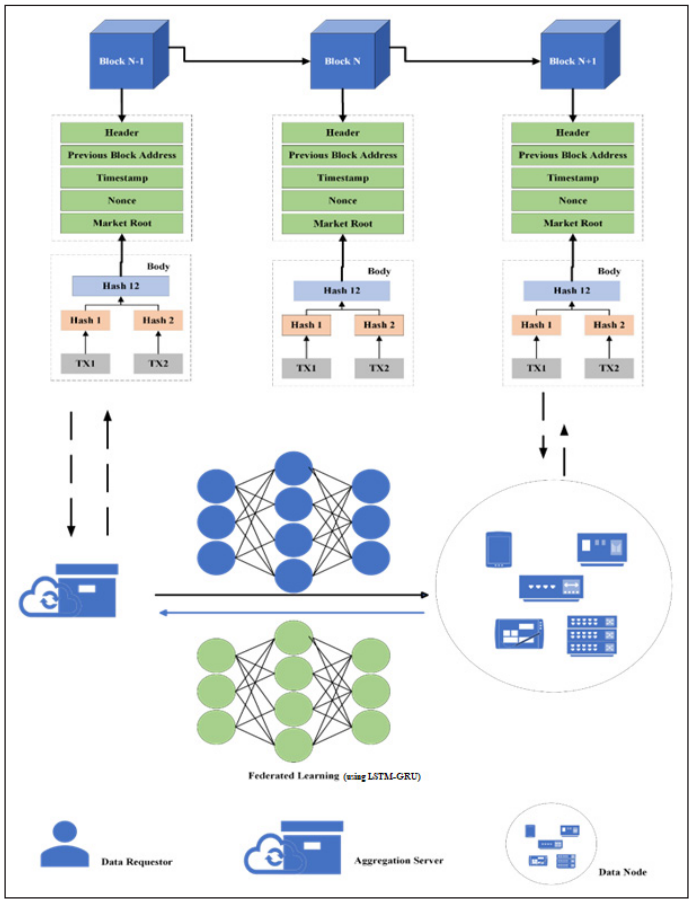

Figure 1 Proposed Model.

the participating nodes reach an agreement, and rewards are allocated based on their contributions. Additionally, data contributors register their data sharing request task on the alliance chain, further promoting the consensus process of the blockchain.

Blockchain (BC) serves as a distributed, open, and decentralized ledger that enables secure storage and transmission of data to cloud servers (Singh et al. 2022). Each block in the blockchain contains transaction information, including timestamps and hash values of previous and succeeding blocks. The cryptographic properties of the blockchain ensure the immutability of data, making it tamper-proof. This decentralized and trustworthy nature of blockchain technology facilitates the distribution of information in a secure and shared manner.

Federated learning, as part of this system model, addresses privacy preservation. It ensures that data remains on local nodes or devices instead of being transferred to a central server. By performing model training locally on each node, federated learning significantly reduces the risk of exposing sensitive information during data transfer. Instead of sharing raw data, only model updates are exchanged, thereby protecting individual data privacy. To further enhance privacy, differential privacy techniques can be applied. These techniques introduce noise or perturbations to the shared model updates, making it challenging to infer or reconstruct individual data.

To confirm secure data transmission during the exchange of model updates in federated learning, encryption and secure communication protocols are used. These measures preserve the transmission from unauthorized access or tampering, minimizing the risks of data leaks and malicious attacks. By leveraging encryption and secure protocols, the confidentiality and integrity of the shared data and model updates are maintained throughout the data sharing process. Furthermore, models that have been trained globally through federated learning can be packaged and recorded on the blockchain. This allows for transparent verification and auditing of the training process. By recording the models on the blockchain, their integrity and authenticity can be ensured, adding an extra layer of trust and accountability to the system.

3.1 LSTM-GRU ARCHITECTURE

In this study, long short-term memory (LSTM) and gated recurrent unit (GRU) approaches are combined to employ predictive analytics. The LSTM-GRU architecture itself does not directly contribute to privacy preservation or secure data transfer. Instead, it is a model architecture commonly used in federated learning for its ability to extract useful patterns. Each of the six hidden levels in the suggested architecture contains 256 hidden units. While the other three are made up of GRU units, three of these hidden layers are made up of LSTM units. Leaky ReLU, a popular non-linear activation function is the activation function employed in each of these hidden layers. The output layer uses one dense layer and one unit with a linear activation function. The output from the dense layer is conveyed to the output layer by lowering the output dimension from the preceding levels. The output values are not constrained to a particular range because of the linear activation function in the output layer, which is useful in some applications.

3.1.1 Gated Recurrent Unit (GRU)

The GRU was designed to address the issue of bursting or disappearing gradients. It is an enhanced version of the LSTM model that also uses gate structures to regulate information flow. It is significant to note that GRU lacks an output gate, making all data accessible to anybody. The input and forget gates are combined in the LSTM, whereas the reset and update gates are the only two gates in GRUs. GRUs perform better because they have fewer parameters and a more straightforward structure. The following equations represent GRU reset and update gates:

$$r_t = \sigma\left(W_r\left[h_{\{t-1\}}, x_t\right] + U_r h_{\{t-1\}} + b_r\right) \tag{1}$$

$$z_t = \sigma\left(W_z\left[h_{\{t-1\}}, x_t\right] + U_z h_{\{t-1\}} + b_z\right) \tag{2}$$

$$\widehat{(h_t)} = \tanh\left(W_h\left[r_t * h_{\{t-1\}}, x_t\right] + b_h\right) \tag{3}$$

$$h_t = (1 - z_t) * h_{\{t-1\}} + z_t * \widehat{(h_t)} \tag{4}$$

Where w

r_t : reset gate at time step t.

z_t : update gate at time step t.

$h_{\{t-1\}}$: hidden state at time step t–1.

x_t : input at time step t.

w_r, w_z, w_h : weight matrices for the reset, update, and candidate hidden state calculation.

u_r, u_z, u_h : weight matrices applied to the hidden state for the reset, update, and candidate hidden state calculation.

b_r, b_z, b_h : biases for the reset, update, and candidate hidden state calculation.

σ : sigmoid activation function.

$\widehat{(h_t)}$: candidate hidden state at time step t.

W : weight matrix for the interpolation calculation.

b : bias label for the interpolation calculation.

h_t : hidden state at time step t.

3.1.2 Long Short-Term Memory (LSTM)

An LSTM, a kind of recurrent neural network, has three gates. The forget, input, and output gates are some of them. The LSTM's vanishing gradient technique causes the gradient in conventional RNNs to disappear (Qiang, Liu & Jin 2021). The forget gate is a crucial factor in determining whether to preserve or delete previously learned information. It assesses the significance of the data from the cell state of the previous time step and decides whether to keep or delete it. The Oblivion Gate's mathematical formula is as follows:

$$p_t = \sigma\left(W_{p\left[h_{\{t-1\}}, x_t\right]} + b_p\right) \tag{5}$$

The performance of the input gate can be calculated through the following formula given below:

$$q_t = \sigma\left(W_{q\left[h_{\{t-1\}}, x_t\right]} + b_q\right) \tag{6}$$

$$v_t = \tanh\left(W_{v\left[h_{\{t-1\}}, x_t\right]} + b_v\right) \tag{7}$$

The computational equation for the output gate is as:

$$f_t = \sigma\left(W_{f\left[h_{\{t-1\}}, x_t\right]} + b_f\right) \tag{8}$$

$$h_t = f_t * v_t + \left(1 - f_t\right) * h_{\{t-1\}} \tag{9}$$

Where

p_t : forget gate activation vector at time step t

x_t : input at time step t

b_p : bias vector for the forget gate calculation

w_p : weight matrix for the forget gate calculation

$h_{\{t-1\}}$: previous hidden state at time step t–1

q_t : input gate activation vector at time step t

v_t : vector of new candidate cell state values at time step t

b_q, b_v : bias vectors for the input gate and candidate cell state calculation

w_q, w_v : weight matrices for the input gate and candidate cell state calculation

f_t : output gate activation vector at time step t

h_t : output at time step t

The federated learning and the LSTM-GRU architecture contribute to privacy preservation and secure data transfer. Differential privacy protection (Wu et al. 2022) and mutual supervision mechanisms have been implemented to mitigate risks associated with data leaks and malicious attacks during data sharing. Federated learning ensures that data remains on the local nodes, and only the model updates are exchanged, reducing the risk of exposing sensitive information. The LSTM-GRU architecture, with its gated structures and secure data transmission protocols, further enhances the privacy and security of the learning process.

4 IMPLEMENTATION

An experiment was conducted on the NSL-KDD dataset to test the effectiveness of the proposed network model. Experiments were conducted to determine the optimal sizes of the training and test sets based on the randomization of the dataset. A performance assessment was then conducted on the trained model using the test set.

4.1 DATASET

Intrusion detection models are commonly tested using the NSL-KDD 2015 dataset. There are 12,5973 samples in this dataset, which are divided into normal samples and anomalous samples. The dataset contains 41 different characteristics that are used to describe the samples. Of the total samples in the dataset, 67,343 are classified as normal and 58,630 are classified as anomalous.

4.2 CONFIGURATION SETUP

Local training in the client utilized PyTorch (Paszke et al. 2019) for implementing the deep learning (DL) algorithm. To enable the development of the federated learning (FL) algorithm, PySyft (Ryffel et al. 2018), a Python extension library compatible with major DL frameworks like PyTorch and TensorFlow (Géron et al. 2019), was employed. PySyft provides the necessary requirements for FL algorithms and facilitates the development of secure and private DL algorithms. The implementation was conducted on the Google Colab platform (Google Colab 2023) with GPU acceleration for efficient processing.

Regarding data preprocessing, the data was initially cleaned and then normalized using StandardScaler. The train set was assigned 80% of the data, while the test set received 20%, following a widely used method. Adaptive FL algorithm optimization was performed using the SGD (Le et al. 2011) optimizer. In the proposed LSTM-GRU model, hyperparameters such as 200 epochs, 0.1 learning rate, and 128 batch sizes were set using a checkpoint to identify the most effective values.

4.3 EVALUATION METRICS

To evaluate the prediction accuracy, five distinct regression evaluation metrics were utilized: Sensitivity (TPR), Specificity (SPC), Precision (PPV), Accuracy (ACC), F1 Score (F1), and Matthews Correlation Coefficient (MCC).

$$TPR = TP / (TP + FN) \tag{10}$$

$$SPC = TN / (FP + TN) \tag{11}$$

$$PPV = TP / (TP + FP) \tag{12}$$

$$ACC = (TP + TN) / (TP + TN + FP + FN) \tag{13}$$

$$F1 = 2TP / (2TP + FP + FN) \tag{14}$$

$$MC = \frac{(TP*TN - FP*FN)}{\text{sqrt}((TP + FP)(TP + FN)(TN + FP)*(TN + FN))} \tag{15}$$

5. RESULTS AND DISCUSSIONS

This work used the NSL-KDD 2015 dataset to validate the proposed model's performance in blockchain based privacy prediction. Based on 20% ratios of testing sets, there were 11,726 anomalous samples and 13,468 normal samples and the proposed model produced results as presented in the Table 1.

Table 1 Model Assessment.

SENSITIVITY (TPR)	SPECIFICITY (SPC)	PRECISION (PPV)	ACCURACY (ACC)	F1 SCORE (F1)	MATTHEWS CORRELATION COEFFICIENT (MCC)
0.9872	0.9927	0.9916	0.9901	0.9894	0.9802

Table 1. shows the evaluation metrics of a proposed blockchain-based model that uses the NSL-KDD 2015 dataset and LSTM-GRU architecture. Evaluation metrics were computed for the testing data, which comprised 80% training data and 20% testing data. In addition to Sensitivity, Specificity, Precision, Accuracy, and Matthews Correlation Coefficient (MCC), we measured Sensitivity (TPR), Specificity (SPC), Precision (PPV), Accuracy (ACC), and F1 Score (F1). As a ratio of total positive samples to the number of true positives, sensitivity (TPR) measures the true positive rate of a model. It was found that 98.72% of anomalous samples in the testing data were accurately identified by the proposed model, whose sensitivity is 0.9872. Specificity (SPC) measures the true negative rate of the model and is calculated as the ratio of true negatives to the total number of actual negative samples. The specificity of the proposed model is 0.9927, indicating that it accurately identified 99.27% of the normal samples in the testing data. PPV measures the ratio of true positives to predicted positives, which is the positive predictive value of the model. The precision of the proposed model is 0.9916, which indicates that when it predicted an anomalous sample, it was correct 99.16% of the time. As the ratio of correctly classified samples to the total number of samples, accuracy (ACC) measures the overall correctness of the model. As a result of the proposed model's accuracy, 99.01% of samples were correctly classified. This metric provides a balanced measure by combining precision and recall with the F1 Score (F1). This model scores 0.9894 on the F1 test, which indicates a good balance between precision and recall. The Matthews Correlation Coefficient (MCC) measures how closely classes are related when taken into account the imbalance in class distributions. In this case, the MCC is 0.9802, which indicates a strong correlation between the actual and predicted classes.

Table 2 displays the confusion matrix generated from the classification of the proposed technique utilizing the NSL-KDD intrusion dataset. The confusion matrix indicates that out of the total samples, the model classified 11,628 samples as anomalies and 13,317 samples as normals. The table also provides additional information on the true positive, false positive, true negative, and false negative values of the classification. Specifically, the model correctly classified 11,628 attack samples as attacks (TP), but incorrectly classified 98 attack samples as normals (FN). Additionally, the model correctly classified 13,317 normal samples as normals (TN), but incorrectly classified 151 normal samples as attacks (FP).

Table 2 Result of Confusion matrix.

ACTUAL	PREDICTED	
	Attack	Normal
Attack	11628	98
Normal	151	13317

Apart from demonstrating the classification accuracy results, the performance of the proposed technique was evaluated using the ROC Curve. Figure 2 presents a visual representation of the classification accuracy results through a ROC curve, which effectively depicts the correlation between the amount of training data and the performance of the technique.

Table 3 highlights and compares the performance of the proposed model with other recent models in terms of accuracy. It demonstrates that the Privacy-Preserving Secure Framework using LSTM-GRU achieved a higher accuracy rate of 99.01% compared to the other models.

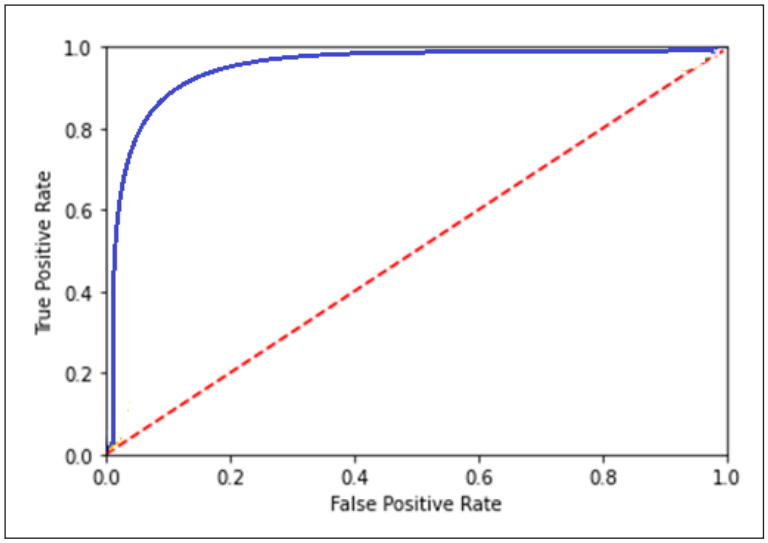

Figure 2 ROC Curve.

Table 3 Comparative analysis.

REFERENCE	MODEL	ACCURACY (%)
K. Pradeep Mohan Kumar et al (2022)	PPSF-BODL	97.46
Alatawi, Mohammed Naif, et al (2023)	PSO-GA followed by ELM-BA	96.04
Proposed	Privacy-Preserving Secure Framework using LSTM-GRU	99.01

6. CONCLUSION

In this study, a novel framework is presented that utilizes blockchain technology in collaboration with multiple contributors, incorporating federated learning for secure and privacy-preserving model training without centralized data storage. Through this approach, models can be trained simultaneously by multiple parties while retaining their local privacy. The framework utilizes federated learning to improve the accuracy of the results, boost model performance, and enhance overall model performance. Transactions are stored in a decentralized, distributed digital ledger, and data privacy and security are ensured through various methods. The LSTM-GRU model, included in the framework, facilitates primary data collection using sensing tools. Experimental results demonstrate the superiority of this approach over existing methods, with an accuracy of 99.01%. The research focused on the NSL-KDD dataset, a widely accepted benchmark for evaluating intrusion detection models, due to its suitable size and characteristics for initial experimentation and proof-of-concept studies. However, it is important to acknowledge that latency in network communication can impact the efficiency of the training process in federated learning. Additionally, blockchain technology may face scalability challenges, which need to be addressed. Future work aims to explore additional datasets with larger sample sizes and a wider range of intrusion scenarios, employing advanced deep learning algorithms to further enhance detection results.

DATA ACCESSIBILITY STATEMENT

Available upon reasonable request.

FUNDING INFORMATION

This work was supported by the Science and Technology Planning Project of He'nan Province, China (Grant No. 222102320343, 222102320351).

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Junzhou Li, Qianhui Sun and Feixian Sun worked together to carry out the research, write the manuscript and draw the figures.

AUTHOR AFFILIATIONS

Junzhou Li orcid.org/0009-0009-1172-370X
Information Management Center of Kaifeng University, Kaifeng 462000, CN
Qianhui Sun orcid.org/0009-0009-3238-3847
International Business School of Henan University, Zhengzhou 450046, CN
Feixian Sun orcid.org/0009-0002-8049-9317
School of Electronics and Internet of Things of Henan Polytechnic, Zhengzhou 450046, CN

REFERENCES

Alatawi, MN, et al. 2023. Cyber security against intrusion detection using ensemble-based approaches. *Security and Communication Networks*, 2023(SI): 1–7. DOI: https://doi.org/10.1155/2023/8048311

Benhar, H, Idri, A and **Fernández-Alemán, JL.** 2020. Data preprocessing for heart diseaseclassification: A systematic literature review. *Computer Methods and Programs in Biomedicine*, 195: 105635. DOI: https://doi.org/10.1016/j.cmpb.2020.105635

Bos, JW, Lauter, K and **Naehrig, M.** 2014. Private predictive analysis on encrypted medical data. *Journal of Biomedical Informatics*, 50: 234–243. DOI: https://doi.org/10.1016/j.jbi.2014.04.003

Chollet, F. 2017. Xception: Deep learning with depthwise separable convolutions. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1251–1258. DOI: https://doi.org/10.1109/CVPR.2017.195

Choudhury, O, et al. 2020. Anonymizing data for privacy-preserving federated learning [online]. [Preprint]. [Viewed 21 February]. Available from: arXiv:2002.09096

Domadiya, N and **Udai, PR.** 2021. Improving healthcare services using source anonymous scheme with privacy preserving distributed healthcare data collectionmining. *Computing*, 103(1): 155–177. DOI: https://doi.org/10.1007/s00607-020-00847-0

Dwork, C. 2011. *Differential privacy encyclopedia of cryptography and security*. Boston: Springer. pp. 338–340. DOI: https://doi.org/10.1007/978-1-4419-5906-5_752

Géron, A. 2019. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. Sebastopol, CA: O'Reilly Media.

Google Colab. Available at https://colab.research.google.com/ [Last accessed 02 May 2023].

Hamza, R, et al. 2020. A privacy-preserving cryptosystem for IoT E-healthcare. *Information Sciences*, 527: 493–510. DOI: https://doi.org/10.1016/j.ins.2019.01.070

He, K, et al. 2016b. Identity mappings in deep residual networks. In: *European Conference on Computer Vision*, Amsterdam, The Netherlands, October 11–14, 2016, pp. 630–645. DOI: https://doi.org/10.1007/978-3-319-46493-0_38

He, K, Zhang, X, Ren, S and **Sun, J.** 2016a. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778. DOI: https://doi.org/10.1109/CVPR.2016.90

Howard, AG, et al. 2017. *Mobilenets: E'cient convolutional neural networks for mobile vision applications* [online]. [Preprint]. Available from arXiv:1704.04861. DOI: https://doi.org/10.1109/CVPR.2017.243

Huang, G, et al. 2017. Densely connected convolutional networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 21–26 July 2017, pp. 4700–4708. DOI: https://doi.org/10.1109/CVPR.2017.243

Jamil, F and **Kim, D.** 2021. An Ensemble of a prediction and learning mechanism for improving accuracy of anomaly detection in network intrusion environments. *Sustain*, 13(18): 10057. DOI: https://doi.org/10.3390/su131810057

Jena, M, Debaprada, et al. 2021. *Intelligent Data Engineering and Analytics*. Singapore: Springer. pp. 507–514. DOI: https://doi.org/10.1007/978-981-15-5679-1_49

Krizhevsky, A, Sutskever, I and **Hinton, GE.** 2012. Imagenet classi¦cation with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Le, QV, et al. 2011. On optimization methods for deep learning. In: *International Conference on Machine Learning*, Bellevue, Washington, USA, June 28–July 2, 2011, pp. 265–272.

Li, C, Li, G and **Varshney, PK.** 2022. Decentralized federated learning via mutual knowledge transfer. *IEEE Internet of Things Journal*, 9: 1136–1147. DOI: https://doi.org/10.1109/JIOT.2021.3078543

Li, J, et al. 2021. A federated learning based privacy-preserving smart healthcare system. *IEEE Transactions on Industrial Informatics*, 18(3): 2021–2031. DOI: https://doi.org/10.1109/TII.2021.3098010

Li, Z, et al. 2020. CrowdSFL: A secure crowd computing framework based on blockchain and federated learning. *Electronics*, 9(5): 773. DOI: https://doi.org/10.3390/electronics9050773

McMahan, B and **Ramage, D.** 2017. Federated learning: Collaborative machine learning without centralized training data. *Google Research Blog*, [Blog].

Paszke, A, et al. 2019. PyTorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32: 8024–8035.

Pradeep, K, Kumar, M, et al. 2022. Privacy preserving blockchain with optimal deep learning model for smart cities. *CMC*, 73(3). 2 DOI: https://doi.org/10.32604/cmc.2022.030825

Qiang, W, Liu R and **Jin, H.** 2021. Defending CNN against privacy leakage in edge computing via binary neural networks. *Future Generation Computer Systems*, 125: 460–470. DOI: https://doi.org/10.1016/j.future.2021.06.037

Rahman, SA, et al. 2020. A Survey on federated learning: The journey from centralized to distributed on-site learning and beyond. *IEEE Internet of Things Journal*, early access.

Rui, Hu, et al. 2020. Personalized federated learning with differential privacy. *IEEE Internet of Things Journal*, 7(10): 9530–9539. DOI: https://doi.org/10.1109/JIOT.2020.2991416

Ryffel, T, et al. 2018. A generic framework for privacy preserving deep learning [online]. [Preprint] Available from: arXiv:1811.04017.

Saba, T, et al. 2022. Securing the IoT system of smart city against cyber threats using deep learning. *Discrete Dynamics in Nature and Society*, 2022: 1–9. DOI: https://doi.org/10.1155/2022/1241122

Shaham, S, et al. 2021. Privacy preserving location data publishing: A machine learning approach. *IEEE Transactions on Knowledge and Data Engineering*, 33(9): 3270–3283. DOI: https://doi.org/10.1109/TKDE.2020.2964658

Simonyan, K and **Zisserman, A.** 2014. *Very deep convolutional networks for large-scale image recognition* [online]. [Preprint]. Available from: arXiv:1409.1556.

Singh, S, et al. 2022. A framework for privacy-preservation of IoT healthcare data using Federated Learning and blockchain technology. *Future Generation Computer Systems*, 129: 380–388. DOI: https://doi.org/10.1016/j.future.2021.11.028

Srivastava, GP, et al. 2021. PPSF: A privacy-preserving and secure framework using blockchain-based machinelearning for IoT-driven smart cities. *IEEE Transactions on Network Science and Engineering*, 8(3): 2326–2341. DOI: https://doi.org/10.1109/TNSE.2021.3089435

Sun, Z, et al. 2019. Differential privacy for data and model publishing of medical ata. *IEEE Access*, 7: 152103–152114. DOI: https://doi.org/10.1109/ACCESS.2019.2947295

Szegedy, C, et al. 2016. Rethinking the inception architecture for computer vision. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826. DOI: https://doi.org/10.1109/CVPR.2016.308

Tanwar, S, et al. 2020. Machine learning adoption in blockchain-based smart applications: The challenges and a way forward. *IEEE Access*, 8: 474–488. DOI: https://doi.org/10.1109/ACCESS.2019.2961372

Toyoda, K, et al. 2020. Blockchain-enabled federated learning with mechanism design. *IEEE Access*, 8: 219744–219756. DOI: https://doi.org/10.1109/ACCESS.2020.3043037

Wang, S, et al. 2019. Adaptive federated learning in resource constrained edge computing systems, *IEEE Journal on Selected Areas in Communications*, 37(6): 1205–1221. DOI: https://doi.org/10.1109/JSAC.2019.2904348

Weng, J, et al. 2019. DeepChain: Auditable and privacy-preserving deep learning with blockchain-based incentive. *IEEE Transactions on Dependable and Secure Computing*, 18(5): 2438–2455. DOI: https://doi.org/10.1109/TDSC.2019.2952332

Wu, X, et al. 2022. An adaptive federated learning scheme with differential privacy preserving. *Future Generation Computer Systems*, 127: 362–372. DOI: https://doi.org/10.1016/j.future.2021.09.015

Yar, H, et al. 2021. Lung nodule detection and classification using 2D and 3D convolution neural networks (CNNs). *Artificial Intelligence and Internet of Things*, 2021: 365–386. DOI: https://doi.org/10.1201/9781003097204-17

Yin, L, et al. 2021. A privacy-preserving federated learning for multiparty data sharing in social IoTs. *IEEE Transactions on Network Science and Engineering*, 8(3): 2706–2718. DOI: https://doi.org/10.1109/TNSE.2021.3074185

Zhao, Y, et al. 2019. Machine learning based privacy-preserving fair data trading in big data market. *Informing Science*, 478: 449–460. DOI: https://doi.org/10.1109/ACCESS.2019.2909559

Zhao, J, Chen, Y and **Zhang, W.** 2019. Differential privacy preservation in deep learning: Challenges, opportunities and solutions. *IEEE Access*, 7: 48901–48911.

TO CITE THIS ARTICLE:
Li, J, Sun, Q and Sun, F. 2023. Enhancing Privacy-Preserving Intrusion Detection in Blockchain-Based Networks with Deep Learning. *Data Science Journal*, 22: 31, pp. 1–11. DOI: https://doi.org/10.5334/dsj-2023-031

Submitted: 18 May 2023
Accepted: 12 July 2023
Published: 31 August 2023

Data Science Journal is a peer-reviewed open access journal published by Ubiquity Press.