International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Advanced Privacy-Preserving Framework Using Homomorphic Encryption and Adaptive Privacy Parameters for Scalable Big Data Analysis

R. Shanthi[1], M. Dinesh Babu[2], N. Kousika[3], C. Vijayaraj[4], Shruti Bhargava Choubey[5],

S. Sambooranalaxmi[6]

Submitted: 01/11/2023 **Revised**: 20/12/2023 **Accepted**: 02/01/2024

Abstract: As smart cyber-physical systems advance, they generate substantial valuable data from healthcare, smart homes, and vehicles, often containing sensitive information. This data requires sanitization for safe analysis. However, rapid data generation necessitates scalable privacy-preserving methods with high privacy and utility. Balancing privacy and utility pose a common challenge in preserving data privacy. The Advanced Privacy-Preserving Framework ensures secure data preprocessing for scalable big data analysis. It segments raw data, enabling distributed computation while preserving privacy through homomorphic encryption. Within each segment, normalization and scaling maintain accuracy without compromising privacy. Adaptive privacy parameters and encrypted noise perturbation ensure differential privacy and statistical integrity. Aggregated results remain encrypted until decryption under stringent privacy conditions. Secure data release, compliant with privacy regulations, includes protective measures like random swapping or masking. The Enhanced Privacy-Preserving Data Perturbation Algorithm partitions, encrypts, sorts, perturbs, and securely releases datasets based on a specified threshold. These steps ensure robust privacy and secure data release throughout the analysis.

Keywords: *Big data, Internet of Things (IoT), data mining, privacy-preserving, smart cyber-physical systems, sensor data streams*

1. Introduction

The escalating prevalence of Smart Cyber-Physical Systems (SCPS) such as smart healthcare, homes, vehicles, and grid systems arises from rapid technological advancements. These systems gather extensive data pivotal for enhancing efficiency and intelligence across various life activities. However, this data often encases sensitive information, raising substantial privacy concerns [1]. Striking a balance between controlled information release and the necessity for sharing data for analysis, such as machine learning and data mining, poses a significant challenge. Safeguarding the privacy of the immense and swiftly generated data streams produced by SCPS necessitates robust, scalable, and efficient solutions [2].

Privacy-preserving data mining (PPDM) grapples with challenges related to the sheer volume and pace of data flow [3]. Two primary approaches in PPDM are encryption and data perturbation. While encryption promises robust security, its computational intricacies might prove impractical for managing vast SCPS-generated data. Data perturbation, encompassing techniques like noise addition and randomization, presents a less intricate yet potentially utility-compromising alternative [4]. Earlier privacy models like l-diversity and k-anonymity have displayed vulnerabilities to specific attacks. Differential privacy (DP) mitigates these vulnerabilities by minimizing the risk of private data leakage [5]. However, DP encounters limitations concerning small databases and struggles with extremely large or continually expanding databases, leading to potential information leaks [6].

Existing perturbation mechanisms often overlook the trade-off between data utility and privacy enhancement [7]. The inefficient processing of high data volumes and streams further complicates privacy preservation within SCPS settings. Addressing these complexities necessitates novel approaches specifically tailored to preserving privacy in data generated by SCPS. The objectives of the proposed work are:

- Develop an advanced framework for big data analysis that effectively balances privacy preservation with data utility.

[1] *Assistant Professor , Department of computer application, B.S Abdur Rahman Crescent Institute of Science and Technology, Vandalur, Tamil Nadu 600048, India. Email:shanthirajaji@gmail.com*
[2] *Professor, Department of Mechanical Engineering, Rajalakshmi Institute of Technology, Chembarambakkam, Chennai ,*
Tamil Nadu 600124. Email:dinesh198014@yahoo.com
[3] *Assistant Professor , Department of Computer Science and Engineering, Sri Krishna College of Engineering and Technology, Kuniyamuthur, Tamil Nadu 641008, India.. Emil: kousika@skcet.ac.in*
[4] *Assistant Professor, Department of Computer Science and Engineering, Kommuri Pratap Reddy Institute of Technology, Medchal, Hyderabad, Telangana State-501301. Email: vijayrj389@gmail.com*
[5] *Associate Professor, Department of Electronics and Communication Engineering, Sreenidhi Institute of Science and Technology, Hyderabad-501301.Telangana,India.Email: shrutibhargava@sreenidhi.edu.in*
[6] *Assistant Professor, Department of Electronics and Communication Engineering, P. S. R Engineering College, Sevalpatti, Sivakasi-26140.Tamil Nadu. India. Email: sambooagnee@gmail.com*

- Create a three-step approach to retain the spatial layout of data while ensuring privacy through sensitivity assessment, polynomial interpolation, and perturbation.
- Implement a robust preprocessing method utilizing encryption, normalization, and adaptive privacy parameters to maintain differential privacy in analysis while securely releasing insights.
- Apply Lagrange interpolation with Gaussian noise to protect privacy while reconstructing or perturbing data within differential privacy principles.
- Develop an enhanced data perturbation algorithm ensuring differential privacy at each step of data partitioning, encryption, noise addition, and secure release, fostering data usability and confidentiality in big data analytics.

2. Literature review

Smart cyber-physical systems (SCPS) encompassing various domains like healthcare, smart cities, and vehicles gather substantial data for analysis. However, security and privacy concerns amid this data surge remain significant [8]. Numerous studies underscore the importance of securing SCPS, especially due to the extensive usage of personally identifiable information (PII). Privacy in SCPS is addressed through mechanisms like authentication, attribute-based encryption, and access control protocols [9].

Data privacy approaches in SCPS mainly revolve around data perturbation and encryption. While encryption offers robust security, its computational complexity limits its application in resource-constrained devices [10]. Perturbation methods, including input and output perturbation, present viable solutions. However, existing models like k-anonymity and l-diversity exhibit vulnerabilities [11]. Differential privacy (DP) emerges as a powerful model, yet faces challenges in real-valued numerical data and scalability for high-dimensional datasets and streams [12].

Key perturbation methods, like additive and multiplicative perturbation, encounter reconstruction attacks, especially with high-dimensional data. Recent methods attempt to address these issues but encounter limitations in utility, scalability, and privacy protection [13]. Existing mechanisms struggle to balance privacy, utility, and scalability in the dynamic SCPS environment. The review exposes the pressing need for an efficient, scalable, and privacy-preserving solution tailored specifically for SCPS-generated data [14]. The major research gaps and challenges of existing systems are as follows:

- Balancing privacy, utility, and scalability in SCPS-generated data.
- Improving DP mechanisms for real-valued numerical data and data streams.
- Addressing reconstruction attacks in perturbation methods, especially with high-dimensional data.
- Developing efficient, scalable, and robust privacy-preserving solutions for SCPS in resource-constrained environments.

3. Proposed work

The advanced privacy-preserving framework tailored for scalable big data analysis emphasizes the delicate equilibrium between privacy and utility, acknowledging their potential interplay. In the context of data mining tasks like classification and clustering, the spatial arrangement of data profoundly influences outcomes. However, privacy measures, such as randomization, might disrupt this spatial layout, impacting utility. Conversely, prioritizing utility could compromise privacy. To navigate this challenge, the framework orchestrates a three-step process: (1) assessing dataset sensitivity to gauge the necessary random noise for effective privacy, (2) utilizing polynomial interpolation with noise caliberated to approximate a noisy function representing the original data, and (3) leveraging this function to generate perturbed data. This approach ensures privacy preservation while upholding the spatial structure of the initial dataset. The methodology involves polynomial interpolation in addition to the calibrated noise addition in alignment with the principles of differential privacy.

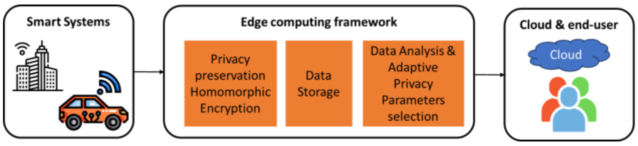
Smart Systems
Edge computing framework
Privacy preservation Homomorphic Encryption
Data Storage
Data Analysis & Adaptive Privacy Parameters selection
Cloud & end-user
Cloud

Fig. 1 Architecture of the proposed model

Figure 1 illustrates the incorporation of the proposed algorithm into the broader data flow of SCPS (Smart Cyber-Physical Systems). The depiction highlights that the perturbed data, resulting from the algorithm, directly originates from SCPS. This indicates that the information stored within the system's storage module has already undergone the privacy preservation process facilitated by the algorithm, thereby ensuring that the stored data no longer retains any original, unaltered information.

3.1 Data Pre-processing

In the Advanced Privacy-Preserving Framework for Scalable Big Data Analysis, data preprocessing involves several essential steps. Firstly, the raw dataset is partitioned into smaller batches to facilitate distributed computation and maintain privacy boundaries per segment. Following this, homomorphic encryption techniques are applied to encrypt the partitions, allowing computations on encrypted data, thereby preserving privacy throughout the analysis process. The data within each partition is normalized and scaled for uniformity, aiding accurate computations while maintaining privacy. Adaptive privacy parameters are then assigned based on data sensitivity or analysis needs, and noise is added to the encrypted data to ensure differential privacy without compromising statistical properties. Secure computation occurs on the encrypted and perturbed data, safeguarding underlying data privacy. Subsequently, aggregated results are obtained from encrypted computations while maintaining privacy, and decryption occurs only when strictly necessary and under stringent privacy-preserving conditions. Finally, processed insights or data are securely released, adhering to privacy regulations and employing mechanisms like random swapping or masking for additional sensitive information protection.

3.2 Mathematical model

The Lagrange interpolation technique is applied to approximate or reconstruct datasets while maintaining privacy. Lagrange interpolation constructs a polynomial that passes through a given set of data points. In the context of your framework, this can be applied to the perturbation or processing of data while preserving the privacy aspects using homomorphic encryption and adaptive privacy parameters. However, the direct application of Lagrange interpolation within the encryption or perturbation process might not be straightforward, as its primary purpose is to interpolate or approximate functions using known data points.

The equations supporting Lagrange interpolation involve constructing a polynomial that fits a set of data points (x_i, y_i). The general form of the Lagrange interpolation polynomial of degree n for a set of distinct points is given by:

$$P(x) = \sum_{i=0}^{n} y_i \cdot L_i(x) \tag{1}$$

Here, n is the degree of the polynomial, x_i and y_i are the given data points, $L_i(x)$ are the Lagrange basis polynomials defined as:

$$L_i(x) = \prod_{j=0, j\neq i}^{n} \frac{x-x_j}{x_i-x_j} \tag{2}$$

Gaussian noise with a sensitivity of 1 is introduced into the root mean square error (RMSE) minimization process. Gaussian noise is added in a calibrated manner to derive values for a1, a2, a3, and a4, in the interpolation process.

$$CA = B \tag{3}$$

Here, C represents the coefficient matrix derived from factorized expressions. A is the coefficient vector obtained from matrix M. B is the constant vector obtained from the factorized expressions. The matrix C is represented as

$$C = \begin{bmatrix} m_{11} & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{24} \\ m_{31} & m_{32} & m_{33} & m_{34} \\ m_{41} & m_{42} & m_{43} & m_{44} \end{bmatrix}$$

(4)

Where m_{ij} denotes coefficients. The vectors A and B are:

$$A = [a_1 \quad a_2 \quad a_3 \quad a_4]^T$$

(5)

$$B = [b_1 \quad b_2 \quad b_3 \quad b_4]^T$$

(6)

Where a_1, a_2, a_3, a_4 represent the coefficients to approximate data, and b_1, b_2, b_3, b_4 denote constants in the interpolation process.

As the input dataset is normalized within the range of 0 and 1, Gaussian noise ensures a randomized error while adhering to this specified range. In this method, the position of the noise (i.e., where the noise is centered) is adjusted to align with the aim of maintaining local minima of RMSE around 0. The process involves the formulation of a linear system, represented by Equation 8, wherein C denotes the coefficient matrix derived from factorized expressions, A denotes the coefficient vector from matrix M, and B signifies the constant vector from the factorized expressions. Solving this linear system (formed utilizing Equations 4, 5, and 6) allows the acquisition of noisy values for a_1, a_2, a_3, and a_4, thereby approximating the input data series using a stochastic function. As Gaussian noise is introduced in a randomized manner calibrated using a user-defined ε value, the outcomes differ with each calculation, contributing to the randomized interpolation process while ensuring privacy through differential privacy measures.

3.3 Algorithmic framework

Algorithm: Enhanced Privacy-Preserving Data Perturbation Algorithm

```
1: Divide D into data partitions ($w_i$) of size $w_s$
2: x = [1, . . . , $w_s$]
3: normalize x within the bounds of [0, 1]
4: for each $w_i$ do
5:   rep = rep + 1
6:   Dp = []  # empty matrix
7:   # Apply homomorphic encryption to the data partitions
8:   for each attribute, ai in $w_i$ do
9:   sai = sort(ai)  # sorted in ascending order
```

```
        generate M
        generate B
        $A = B \ast M-1$
        # Use A = [a1, a2, a3, a4] to generate perturbed data (api) using ˆ f(x)
        # Adjust ε dynamically based on sensitivity of the attribute or window
        normalize api within the bounds of [0, 1] to generate aNi
        # Apply adaptive privacy parameter selection based on sensitivity/utility
        # Normalize aNi within the bounds of [min(ai), max(ai)]
        # Resort aNi to the original order of ai to generate aoi
    end for
    merge all aoi to generate wpi
    Dp = merge(Dp, wpi)
    if rep == t then
        # Apply necessary privacy-enhancing operations (random swapping, etc.)
        release Dp  # Perform secure release with homomorphic encryption
        rep = 0
    end if
end for
if t == -1 then
    # Apply necessary privacy-enhancing operations (random swapping, etc.)
    release Dp  # Perform secure release with homomorphic encryption
    return Dp
end if
End Algorithm
```

The enhanced privacy-preserving data perturbation algorithm begins by dividing the input dataset 'D' into smaller data partitions of a specified size ('ws'). It initializes parameters and normalizes a sequence 'x' within the range [0, 1]. For each data partition 'wi', the algorithm performs several operations. It involves applying homomorphic encryption to the partitions, sorting attributes within each partition in ascending order, generating matrices based on specific equations, and using them to perturb the data ('api'). Furthermore, the algorithm dynamically adjusts the privacy parameter 'ε' based on attribute or window sensitivity, then normalizing and resorting the perturbed data. These steps ensure that the perturbed data maintains its original order and falls within specific value ranges. After processing all partitions, it merges the perturbed data to generate 'Dp'. The algorithm includes conditions to release 'Dp' securely with privacy-enhancing operations like random swapping, contingent on the repetition count reaching a threshold ('t'). Finally, the algorithm returns the perturbed dataset 'Dp' after ensuring privacy through necessary operations, including secure release with homomorphic encryption, contingent on the value of 't'.

4. Results and discussion

4.1 Experimental Setup

Tests were conducted during the experimental phase on a Windows 10 Professional Edition system (64-bit) that had an Intel i5-6200U CPU (6th generation) with a dual-core configuration capable of handling four logical threads. The CPU had a speed range of 2.3 GHz to 2.8 GHz using turbo boost and was supported by an 8192 MB RAM capacity. The proposed algorithm's scalability was evaluated by conducting experiments on Amazon Web Services (AWS) using a cloud-based infrastructure. An EC2 instance with a configuration of 32 vCPUs, 128GB RAM, and an EBS-backed storage system optimized for high-performance computing was used. The algorithm was implemented using Python 3.9, and data classification experiments were performed using TensorFlow and Scikit-learn libraries, which are renowned for their efficacy in machine learning and data mining tasks.

4.2 Dataset description

A concise portrayal of the datasets is outlined in Table 1. These datasets exhibit diverse dimensions, ranging from small to exceedingly large, deliberately chosen to comprehensively evaluate the proposed algorithm's performance under varying scenarios. Notably, the current iteration of the proposed algorithm focuses solely on perturbing numerical data. Therefore, datasets chosen for evaluation exclusively contain numerical attributes, except for the class attribute, which includes non-numerical data.

Dataset Name	Number of Records	Number of Attributes	Number of Classes/Targets
Wine Quality	1599	12	2 (Red/White)
Iris Dataset	150	4	3 (Iris Species)
House Prices	1460	81	Regression (Price)

Dataset			
Letter Recognition	20000	17	26 (Alphabets)
Diabetes Dataset	768	8	2 (Diabetic/Non-diabetic)
HIGGS	11000000	28	2 (Yes/No)

4.3 Performance metrics and analysis

The analysis involves evaluating the outcomes derived from the proposed model against Additive Perturbation (AP), Hybrid Perturbation (HP), Geometric Perturbation (GP), Random Projection (RP), and Data Condensation (DC). When comparing performance with the proposed model, the assessment was conducted using AP, HP, GP, and RP on static datasets, whereas DC was employed specifically with data streams. These mechanisms were chosen due to their multidimensional perturbation nature, aligning closely with the technique applied in the linear system of the proposed model.

One of the main attributes of the proposed model lies in its capability to perturb a dataset while retaining the original shape of its data distribution. The model was executed on the same data series to observe the impact of randomization in two separate instances of perturbation. This experiment aims to ensure that the proposed model does not generate similar perturbed data when applied with the same ε value to identical data on different occasions. This feature plays a pivotal role in preventing privacy breaches caused by data linkage attacks exploiting multiple data releases. As shown in Figure 2, the proposed model generates two distinct randomized data series in two different applications while maintaining the structure of the original data series. The plot visualizes the data generated under ε values of 1 and 0.1, illustrating the substantial effect of increased randomization at a stricter privacy budget (ε of 0.1).

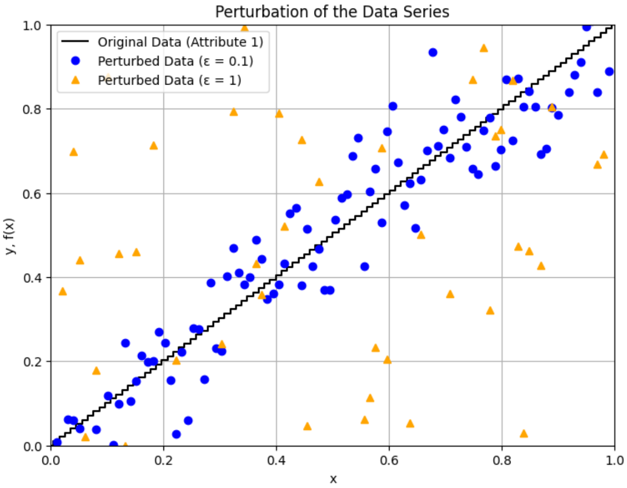

Fig. 2. Effect of perturbation

A decline in utility is evident with reduced ε, indicating higher randomization. Figure 3 illustrates the relationship between classification accuracy and increasing ε, showcasing an upsurge in accuracy as ε grows. Meanwhile, Figure 4 portrays a consistent trend of heightened utility (classification accuracy) with rising ε. The choice of an appropriate ε depends on specific application needs; smaller ε values cater to increased privacy requirements, while larger ε values enhance utility. Notably, double-digit ε values do not offer substantial privacy benefits. Emphasizing the preservation of the original data distribution, an ε range of 0.4 to 3 is recommended to prevent unexpected privacy vulnerabilities. The study underscores the superior privacy and utility of the proposed model compared to similar methods within an ε budget of 1.

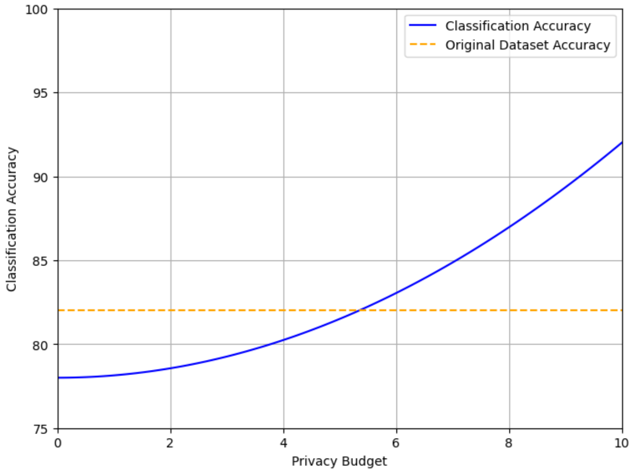

Fig. 3. Classification Accuracy vs. Window Size (ws) at 10,000 Tuples

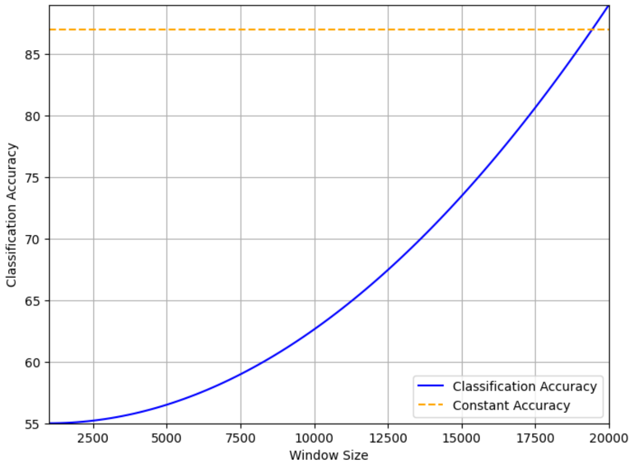

Fig. 4. Classification Accuracy vs. Privacy budget (ε) at 1

Figure 5 displays the time consumption plots of various methods overlaid on a single graph. Notably, the curves representing the proposed model align almost parallel to the x-axis, indicating significantly lower time consumption compared to the other methods.

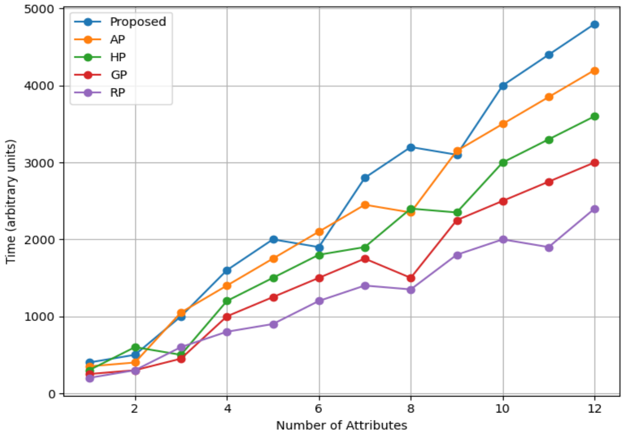

Fig. 5. Time consumption vs. number of attributes

4. Conclusion and Future Scope

The privacy-preserving big data analysis framework achieves a balance between privacy and utility through a three-step process: assessing dataset sensitivity, employing polynomial interpolation with calibrated noise, and generating perturbed data while preserving the original spatial structure. The experiments conducted demonstrate the model's capability to safeguard privacy, maintain data distribution shapes, and mitigate potential privacy breaches. An advised ε range of 0.4 to 3 offers a stronger defense against unforeseen privacy vulnerabilities, exhibiting an improved equilibrium between privacy and utility compared to existing methods within an ε budget of 1. The model's efficiency and effectiveness in ensuring robust privacy preservation while sustaining utility represents a significant stride in privacy-conscious data analysis. Future research endeavors may concentrate on enhancing the model's adaptability to diverse data types and exploring optimization strategies to enhance scalability when handling larger datasets.

References

[1] Sun, Y., Liu, Q., Chen, X., & Du, X. (2020). An adaptive authenticated data structure with privacy-preserving for big data stream in cloud. *IEEE Transactions on Information Forensics and Security*, *15*, 3295-3310.

[2] Chenthara, S., Ahmed, K., Wang, H., & Whittaker, F. (2019). Security and privacy-preserving challenges of e-health solutions in cloud computing. *IEEE access*, *7*, 74361-74382.

[3] Jin, H., Luo, Y., Li, P., & Mathew, J. (2019). A review of secure and privacy-preserving medical data sharing. *IEEE Access*, *7*, 61656-61669.

[4] Yang, Y., Zheng, X., Guo, W., Liu, X., & Chang, V. (2019). Privacy-preserving smart IoT-based healthcare big data storage and self-adaptive access control system. *Information Sciences*, *479*, 567-592.

[5] Zheng, X., & Cai, Z. (2020). Privacy-preserved data sharing towards multiple parties in industrial IoTs. *IEEE Journal on Selected Areas in Communications*, *38*(5), 968-979.

[6] Kaissis, G., Ziller, A., Passerat-Palmbach, J., Ryffel, T., Usynin, D., Trask, A., ... & Braren, R. (2021). End-to-end privacy preserving deep learning on multi-institutional medical imaging. *Nature Machine Intelligence*, *3*(6), 473-484.

[7] Kumar, P., Kumar, R., Srivastava, G., Gupta, G. P., Tripathi, R., Gadekallu, T. R., & Xiong, N. N. (2021). PPSF: A privacy-preserving and secure framework using blockchain-based machine-learning for IoT-driven smart cities. *IEEE Transactions on Network Science and Engineering*, *8*(3), 2326-2341.

[8] Arachchige, P. C. M., Bertok, P., Khalil, I., Liu, D., Camtepe, S., & Atiquzzaman, M. (2020). A trustworthy privacy preserving framework for machine learning in industrial IoT systems. *IEEE Transactions on Industrial Informatics*, *16*(9), 6092-6102.

[9] Lin, J. C. W., Srivastava, G., Zhang, Y., Djenouri, Y., & Aloqaily, M. (2020). Privacy-preserving multiobjective sanitization model in 6G IoT environments. *IEEE Internet of Things Journal*, *8*(7), 5340-5349.

[10] Hao, M., Li, H., Xu, G., Liu, S., & Yang, H. (2019, May). Towards efficient and privacy-preserving federated deep learning. In *ICC 2019-2019 IEEE international conference on communications (ICC)* (pp. 1-6). IEEE.

[11] Deebak, B. D., Al-Turjman, F., Aloqaily, M., & Alfandi, O. (2019). An authentic-based privacy preservation protocol for smart e-healthcare systems in IoT. *IEEE Access*, *7*, 135632-135649.

[12] Yu, K., Tan, L., Shang, X., Huang, J., Srivastava, G., & Chatterjee, P. (2020). Efficient and privacy-preserving medical research support platform against COVID-19: a blockchain-based approach. *IEEE consumer electronics magazine*, *10*(2), 111-120.

[13] Zhou, C., Fu, A., Yu, S., Yang, W., Wang, H., & Zhang, Y. (2020). Privacy-preserving federated learning in fog computing. *IEEE Internet of Things Journal*, *7*(11), 10782-10793.

[14] Fang, H., & Qian, Q. (2021). Privacy preserving machine learning with homomorphic encryption and federated learning. *Future Internet*, *13*(4), 94.