



IJRASET

International Journal For Research in
Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** V **Month of publication:** May 2024

DOI: <https://doi.org/10.22214/ijraset.2024.62868>

www.ijraset.com

Call:  08813907089

E-mail ID: ijraset@gmail.com

Effective Deep Learning Technique for Enhanced Data Privacy and Security

Prof. D. M. Kanade¹, Sharanya Datrang², Rutuja Aher³, Nayan Deshmukh⁴, Divya Tambat⁵

K. K. Wagh Institute of Engineering Education and Research Nashik, India

Abstract: *In the past, data privacy and security during analysis were challenging. Sensitive information often remained vulnerable, risking privacy breaches. This research introduces a comprehensive solution to address these challenges. It consists of three main stages: PII detection, differential privacy with Gaussian noise, and homomorphic encryption. It starts with data collection from various sources. What sets the system apart is its ability to safeguard personal data. This research employs PII detection techniques to identify and anonymize sensitive information, preserving privacy without compromising data utility. Next, preprocess the data, enhancing its quality for analysis. Differential privacy is applied, introducing controlled Gaussian noise and aggregating the data to protect individual privacy while enabling meaningful insights. Moreover, This research uses homomorphic encryption, which allows confidential calculations to be performed without revealing sensitive information. This is especially beneficial for securing Indian household data. As move on to data analysis, the research system leverages machine learning and analytical methods to extract insights from the protected data. Finally, the results are visualized and presented in reports, ensuring that the protected data is effectively utilized while respecting privacy and security concerns. In summary, the system provides a comprehensive solution for handling sensitive data, ensuring privacy, and enabling valuable insights to be drawn from the data without compromising individuals privacy and data security. It significantly enhances data privacy and security compared to the past, where these concerns were inadequately addressed.*

Keywords: *Data privacy, security, homomorphic encryption, sensitive data, Differential privacy.*

I. INTRODUCTION

In today's world of vast datasets, balancing the need for insights with protecting privacy is crucial. This study focuses on safeguarding privacy while analyzing household population data in India. As we dig into population dynamics, it's essential to keep sensitive information safe. This research uses two key privacy techniques: adding noise with Differential Privacy, and using Homomorphic Encryption. By combining these methods, we aim to protect identities and sensitive data while keeping the dataset useful for analysis. Our goal is to create a system that balances data utility and privacy, contributing to secure data analysis methods. This study is trying out a new way to keep people's information safe while still analyzing big sets of data. It's like putting three special locks on a treasure chest: one lock checks for personal info, another adds some random noise to the data, and the last one scrambles the data so no one can peek inside. By using all three locks together, we can make sure no one's personal info gets out, while still keeping the data useful for studying. This research is like building a strong shield to protect people's privacy while we learn from the data. It's all about finding a balance between keeping info private and making sure we can still learn useful things from the data.

DATASET: Household population India: The dataset contains demographic information for various geographic entities within India. The dataset includes parameters such as State Code, District Code, Sub District Code, and the corresponding names of India's States, Union Territories, Districts, and Sub-districts. Additionally, it encompasses details on the total, rural, and urban classification, along with the counts of inhabited and uninhabited villages. Furthermore, the dataset records the number of towns, households, and population statistics, including the total population, male population, and female population. Geospatial details, such as the area covered in square kilometers, are also provided, offering insights into the distribution of population density across different regions.

II. METHODOLOGY

A. Differential Privacy

- 1) For adjacent data sets D and D' that differ by at most one data item, an attacker can infer the presence or absence of this data item from the query function f . The differential privacy algorithm can prevent this kind of attack by the randomized algorithm M . Differential privacy techniques can quantify the degree of protection of sensitive information by the randomized algorithm M .

If the randomized algorithm M privacy budget is high, then the probability of an attacker inferring sensitive data is high. The lower the privacy budget, the more rigorous protection can be applied to data privacy. We give the definition and implementations of differential privacy. First, the formal definition of differential privacy is given as follows:

- 2) Definition 1: differential privacy. For a randomized algorithm M with domain D and range S . For any two adjacent input data x and y , M satisfies differential privacy if it holds that

$$P_r[M(x) \in S] < \exp(\epsilon) \cdot P_r[M(y) \in S] + \delta(1)$$

The key technique of differential privacy algorithm is distorting the data according to the privacy budget. Among the implementations. way of differential privacy, they can be categorized in two ways: (1) adding random noise, (2) random responses. The former is dominated by the Laplace mechanism [9] and the exponential mechanism is implemented by random responses. The Gaussian noise mechanism [7] tries to add noise that obeys the Gaussian distribution to the query function according to the privacy budget and global sensitivity of the data set D . The global sensitivity is the maximum of the absolute distance and it is defined as follows:

- Definition 2: Global Sensitivity. The sensitivity of a given query function f is defined by -

$$\Delta f = \max |f(D) - f(D')|$$

B. Homomorphic Encryption

- 1) Homomorphic encryption is a cryptographic technique that enables computations to be performed on encrypted data without decrypting it first. This property is crucial for preserving the privacy of sensitive information while allowing meaningful operations to be conducted. Homomorphic Encryption (HE) is primitive encryption that allows a party to encrypt data and send it to another party that can then perform certain operations on the encrypted version of the data [10]. An encryption system that allows arbitrary calculations to be encoded on encrypted data without decryption or access to any symmetric cryptographic decryption key is known HE [21]. When the account ends, the encrypted version of the result is sent to the first party that can decrypt and get the result in plain text. Homomorphic Encryption scheme (Enc) follows the following equation:

$$\text{Enc}(a) \hat{\wedge} \text{Enc}(b) = \text{Enc}(a \hat{\wedge} b)$$

where $\text{Enc}: X \hat{\wedge} Y$ is a Homomorphic Encryption scheme wherein X is used for a set of messages and Y is used for ciphertext. Furthermore, a and b are messages in X and $\hat{\wedge}$ are linear operations [6].

- 2) Homomorphic encryption methods can be partially divided into, partially homomorphic and fully homomorphic encryption. In a partially homomorphic system, only one type of operation (either addition or multiplication) can be performed on the encrypted data, while fully homomorphic encryption allows both operations. The fundamental idea behind homomorphic encryption lies in transforming plaintext operations into equivalent operations on encrypted data. This property allows computations on the encrypted data to produce results that, when decrypted, correspond to the desired output of the original operations.
- 3) These formulas illustrate the core principles of homomorphic encryption and its ability to perform computations while maintaining the confidentiality of the underlying data. Utilizing these techniques in research can provide a robust framework for secure data processing. After completing the computations on the encrypted data, obtain the encrypted result. In this we can perform operations on the encrypted data itself.
- 4) If applicable we can convert the encrypted result into a format suitable for further analysis or interpretation. Use the private key to decrypt the final result, ensuring that only authorized entities can access the original outcome. Confirm that the decrypted result aligns with the intended output of the computations performed on the encrypted data.

C. Autoencoder

Autoencoder is a fundamental component of our methodology, contributing to the preservation of data privacy and facilitating meaningful feature learning. In our project, we utilized an autoencoder architecture consisting of multiple layers with Rectified Linear Unit (ReLU) activation functions.

The ReLU activation function introduces non-linearity to the model and aids in efficient gradient propagation, enhancing the learning process.

- 1) **Training Process:** The autoencoder was trained using Household dataset, employing an optimizer and a custom loss function tailored to our project’s objectives. Additionally, privacy concerns were addressed by integrating differential privacy techniques into the training procedure. By distorting the data according to the privacy budget, the autoencoder ensured robust protection against potential privacy breaches.
- 2) **Evaluation Metrics:** Evaluation of the autoencoder’s performance was conducted using reconstruction error analysis. This involved quantifying the discrepancy between the input data and the reconstructed output generated by the autoencoder. Reconstruction error analysis provided valuable insights into the effectiveness of the autoencoder in learning meaningful representations of the dataset.
- 3) **Results and Insights:** Through extensive experimentation, we observed promising results from the autoencoder model. Despite encountering challenges such as noise and complexity inherent in real-world datasets, the autoencoder demonstrated robust capabilities in feature learning and data denoising. These insights underscore the significance of the autoencoder in preserving data privacy and facilitating secure data processing.

The first page must contain, in the following sequence:

D. Figures and tables

A description of the program architecture is presented. Subsystem design or Block diagram.

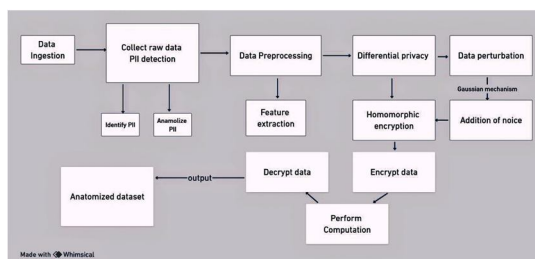


Figure 1: Block Diagram

This research is based on the Household- population dataset. This comprehensive dataset contains a wide range of demographic and population information, making it a valuable resource for various data analysis tasks. However, due to the sensitive nature of the data, privacy preservation and responsible data handling are of utmost importance.

State	District	Col	Sub	District	India/	Stat	Name	Total	Inhabited	Uninhabited	Number	o	Number	o	Person/eq	Male	Popu.	Female	Po Area	Population per sq.
0	0	0	INDIA	INDIA	@	Total		43,324	7,933	2,547,08	1,21479	6,71408	5,91408	3287469	382					
0	0	0	INDIA	INDIA	S	Rural		36,285	43,324	0	1,71408	8,31408	4,31408	4,11408	3103474	279				
0	0	0	INDIA	INDIA	S	Urban		0	0	7,933	8,11407	3,81408	21408	1,81408	102332	3,685				
1	0	0	STATE	JAMMU &	Total			6,337	216	122	2197718	1,31407	6640662	5900640	222236	124				
1	0	0	STATE	JAMMU &	Rural			6,337	216	0	1553433	9108060	4774477	4333583	220990	91				
1	0	0	STATE	JAMMU &	Urban			0	0	122	566285	3833142	1866485	1567057	1245,9	2,755				
1	1	0	DISTRICT	Kopwarea	Total			353	9	10	113929	870354	474190	396164	2379	366				
1	1	0	DISTRICT	Kopwarea	Rural			353	9	0	101900	765625	412038	355847	2331,66	328				
1	1	0	DISTRICT	Kopwarea	Urban			0	0	10	11,999	104729	62,152	42,677	42,34	2212				
1	1	1	SUB-DISTRICT	Kopwarea	Total			118	4	7	63,022	540914	297837	243077	301,04	1,791				
1	1	1	SUB-DISTRICT	Kopwarea	Rural			118	4	0	56,014	465323	252856	212467	275,03	1,692				
1	1	1	SUB-DISTRICT	Kopwarea	Urban			0	0	7	7,008	979660	44,981	30,610	26,91	2,809				
1	1	2	SUB-DISTRICT	Handwara	Total			196	3	1	39,485	269311	141882	127429	291,47	924				
1	1	2	SUB-DISTRICT	Handwara	Rural			196	3	0	37,414	255711	134003	121208	282,97	904				
1	1	2	SUB-DISTRICT	Handwara	Urban			0	0	1	2,071	910806	7,379	6,221	8,5	1,600				
1	1	3	SUB-DISTRICT	Karnah	Total			39	2	2	14,422	910806	34,471	25,658	69,89	860				
1	1	3	SUB-DISTRICT	Karnah	Rural			39	2	0	8,442	910806	24,679	19,612	57,96	709				
1	1	3	SUB-DISTRICT	Karnah	Urban			0	0	2	2,980	910806	9,792	5,746	11,93	1,302				

Figure 2: Database

III. CONCLUSION

In conclusion, removing personally identifiable information (PII) from datasets is not enough to anonymize data. This research work focuses on the privacy and security of sensitive data using the deep learning technique of Differential Privacy. Differential privacy is achieved by adding Gaussian noise to the data. Specifically, it investigates the possibility of performing computations on encrypted data using homomorphic encryption, thus eliminating the need for decryption. This approach aims to maintain the utility of the data while achieving high performance. Additionally, the integration of autoencoder techniques serves to enhance the research’s privacy measures. The autoencoder’s purpose is to learn a compressed representation of the data, ensuring efficient data reconstruction without compromising privacy. By leveraging the autoencoder, our research ensures that sensitive information remains protected throughout the data processing pipeline, ultimately contributing to the overall goal of effective deep learning for data privacy and security.

REFERENCES

- [1] Jiapeng Zhang, Luoyi Fu, Huan Long, Guiâe Meng, Feilong Tang, Xinbing Wang, and Guihai Chen, Member, "Collective De- Anonymization of Social Networks With Op- tional Seeds", IEEE 2021
- [2] Jun Li, Fengshi Zhang, Yonghe Guo, Siyuan Li, Guanjun Wu, Dahui Li, Hongsong Zhu, "A Privacy-Preserving Online Deep Learn- ing Algorithm Based on Differential Pri- vacy", 2023
- [3] Matheus M. Silveira, Ariel L. Portela, Rafael A. Menezes, Michael S. Souza, Danielle S. Silva, Maria C. Mesquita, and Rafael L. Games, "Data Protection based on Search- able Encryption and anonymization Tech- niques", IEEE 2023
- [4] Salaheddine Kabou, Sidi Mohamed Bensli- mane, and Abdelbaset Kabou, "Towardsnew way of minimizing the loss of information quality in the dynamic anonymization", 978- 1-7281-2580-0/20/, IEEE2020
- [5] Peng Chong, "Deep Learning based Sensi- tive Data Detection", IEEE 2022
- [6] Muhammad Imran Tariq, Nisar Ahmed Memon, Shakeel Ahmed, Shahzadi Tayyaba, Muhammad Tahir Mushtaq, Natash Ali Mian, Muhammad Imran, and Muhammad W. Ashraf, "A Review of Deep Learning Se- curity and Privacy Defensive Techniques", Hindawi Mobile Information Systems Vol- ume 2020
- [7] Dong, A. Roth, and W. J. Su, "Gaus- sian differential privacy," arXiv preprint:1905.02383, 2019
- [8] Maryam Archie, Sophie Gershon, Abigail Katcoff, and Aaron Zeng, "Whoâs Watch- ing? De-anonymization of Netflix Reviews using Amazon Reviews"
- [9] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in Theory of cryp- tography conference. Springer, 2006, pp. 265â284
- [10] A. Acar, H. Aksu, A. S. Uluagac, and M. Conti, "A survey on homomorphic encryp- tion schemes," ACM Computing Surveys, vol. 51, no. 4, pp. 1â35, 2018
- [11] S. Halevi, Y. Polyakov, and V. Shoup, "An improved RNS variant of the BFV homo- morphic encryption scheme," in Topics in CryptologyâCT-RSA 2019, pp. 83â105, Springer, Berlin, Germany, 2019
- [12] M. Abadi, A. Chu, I. Goodfellow et al., "Deep learning with differential privacy," in Proceedings of the 2016 ACM SIGSAC Con- ference on Computer and CommunicationsSecurityâCCSâ16, pp. 308â318, Vienna, Austria, 2016.
- [13] N. C. Abay, Y. Zhou, M. Kantarcioglu, B. uraisingham, and L. Sweeney, "Privacy pre- serving synthetic data release using deep learning," in Machine Learning and Knowl- edge Discovery in Databases, pp. 510â526, Springer, Berlin, Germany, 2019
- [14] S. D and K. Karibasappa, "Enhancing data protection in cloud computing using key derivation based on cryptographic tech- nique," in 2021 5th International Conference on Computing Methodologies and Commu- nication (ICCMC), 2021, pp. 291â299.
- [15] C. F. Chiasserini, M. Garetto, and E. Leonardi, "Social network deanonymization under scale-free user relations," IEEE/ACM Trans. Netw., vol. 24, no. 6, pp. 3756â3769, Dec. 2016.
- [16] K. Alrawashdeh and C. Purdy, "Toward an online anomaly intrusion detection system based on deep learning," in Proceedings of the 2016 15th IEEE International Confer- ence on Machine Learning and Applications (ICMLA), pp. 195â 200, Anaheim, CA, USA, December 2016
- [17] A. De Slave, P. Mori, and L. Ricci, "A sur- vey on privacy in decentralized online so- cial networks," Computer Science Review, vol.27,pp.154-176.2018.
- [18] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436â444, 2015
- [19] V. Kotu and B. Deshpande, "Deep learning," in Data Science, pp. 307â342, Elsevier, Am- sterdam, Netherlands, 2019.
- [20] X. Qiu, L. Zhang, Y. Ren, P. N. Suganthan, and G. Amaratunga, "Ensemble deep learn- ing for regression and time series forecast- ing," in Proceedings of the 2014 IEEE Sym- posium on Computational Intelligence in En- semble Learning (CIEL), pp. 1â6, Orlando, FL, USA, December 2014.
- [21] D. Boneh, "Threshold cryptosystems from threshold fully homomorphic encryption," in Advances in Cryptologyâ CRYPTO 2018, pp. 565â596, Springer, Berlin, Germany, 2018.



10.22214/IJRASET



45.98



IMPACT FACTOR:
7.129



IMPACT FACTOR:
7.429



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089  (24*7 Support on Whatsapp)