International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 23s, 2025
https://theaspd.com/index.php

Challenges in Explainable AI and Privacy Preserving Machine Learning

M. Dyna[1], Dr. P. Swetha[2]
[1]Research Scholar, JNTUH, Hyderabad, India
[1]Assistant Professor, CSE Department, Maturi Venkata Subba Rao (MVSR) Engineering College, Hyderabad, India
[2]Professor, CSE Department, JNTUH College of Engineering, Hyderabad, India
dyna_cse@mvsrec.edu.in[1], drpswetha@jntuh.ac.in[2]

Abstract
The primary requirement of several applications using machine learning models is input data. Privacy preserving Machine Learning focuses on preserving the privacy of user's data and the results predicted from the ML model. Membership inference attacks, poisoning attacks, model extraction attacks, reconstruction attacks effect the private information of users and the output. PPML techniques like Homomorphic encryption, Differential Privacy methods, federated learning and trusted execution environments ensure privacy of data. Also, there should be transparency and interpretability of the output generated by the ML models. The need for both explainability in complex models and ensuring privacy is the motivation of this paper.

In this paper the scope of explainable AI (XAI) and Privacy Preserving ML and its significance of ensuring privacy through various Privacy Preserving ML techniques like HE, DP, FL, TEE and transparency through explainable AI is discussed. The objective of this paper is to discuss the various challenges of applying XAI techniques like Shapley Additive Explanations (SHAP) and Local Interpretable Model Agnostic Explanations (LIME) to encrypted data predictions. This survey contributes to the area of Privacy Preserving Machine Learning (PPML) and Privacy Preserving Explainable AI (PPXAI) by providing an insight and discussing various challenges in bridging the gap between PPML and the need for ensuring security, transparency and interpretability.

1. INTRODUCTION

Recent advances in the applications of AI and ML clearly suggests the amount of data that is used. The input data which is used for predictions using ML models needs to be protected from malicious and unidentified users. Simultaneously the predictions made using complex models must be interpretable and explainable. Motivation of this paper arises from various reasons. Transparent Machine Learning Model is required for data analysts to interpret the behavior of the model and its impact which in turn enables them to modify the parameters and improve the accuracy of the model. To build a reliable framework and trustworthy predictions certain guidelines, regulations and ethics like GDPR need to be incorporated. The Objective of Explainable AI is to describe about the predictions and build trust in the models. Privacy preserving techniques like Homomorphic encryption achieves input privacy by encrypting the raw data. Differential privacy technique ensures privacy about the output of a computation by adding noise or by creating synthetic data thus securing the output of the Machine Learning model. Secure multi-party computation enables data protection by processing on encrypted data and splits them to multiple parties such that no single party can retrieve the entire data on their own. Further Explainable AI is needed to explain and assure the stakeholders about the privacy of raw data or the results of the Machine Learning models that use encryption techniques. The challenge here is to explain the end users the predictions made by the model and also the implication of their choice of model which is called the human data interaction (HDI) while providing security.

In this survey we discuss the need of various PPML techniques like Homomorphic Encryption, Differential Privacy, Trusted Execution Environment, Secure Multi-party Computation to ensure privacy and the challenges that arise by using the said techniques. further Explainable AI focuses on HDI principles like legibility, agency and negotiability. Legible explanation is needed to show how the users' data is protected and preserved. Agency refers the right to be forgotten where the Machine Learning Model was developed using wrong data. Negotiability enables the data owners to reevaluate their data sharing decisions. The aim of this paper is to summarize the threats that could occur, methods to preserve privacy and discuss the challenges of Privacy Preserving Machine Learning .The paper further contributes to future research that can bridge the gap between PPML and the need for transparency, interpretability and trustworthiness through Explainable AI.

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 23s, 2025
https://theaspd.com/index.php

2. Background and Foundation

The rapid advancement of Artificial Intelligence (AI) and Machine Learning (ML) has led to their widespread adoption in various domains, including healthcare, finance, and cybersecurity. While these technologies offer significant benefits, they also raise concerns about data privacy and model interpretability. Ensuring privacy while leveraging the power of ML and making AI models more explainable have become critical research areas. This section provides an overview of two fundamental aspects: Privacy-Preserving Machine Learning (PPML) and Explainable AI (XAI).

2.1Privacy Preserving Machine Learning (PPML)

The information that is needed to recognize an individual for example: Aadhar Number, email or a telephone number is called personally identifiable Information (PII). This data is used in many applications like healthcare, financial, social media platforms. In medical data analytics, the patient data is used to predict diseases and to provide improve treatment recommendations. PPML ensures to keep the PII local and shares only the model updates to modify and refine the treatments. With reference to the paper[1](ensuring data Privacy using ML for responsible Data Science) , Quasi identifiers (QI) refers to the inquiry results to recognize an individual for example PIN code, Age. Confidential information data like the Bank account details, live geo area can also be used in financial applications like credit scoring and fraud detection. Sensitive and personal information breach results in Inference attacks.

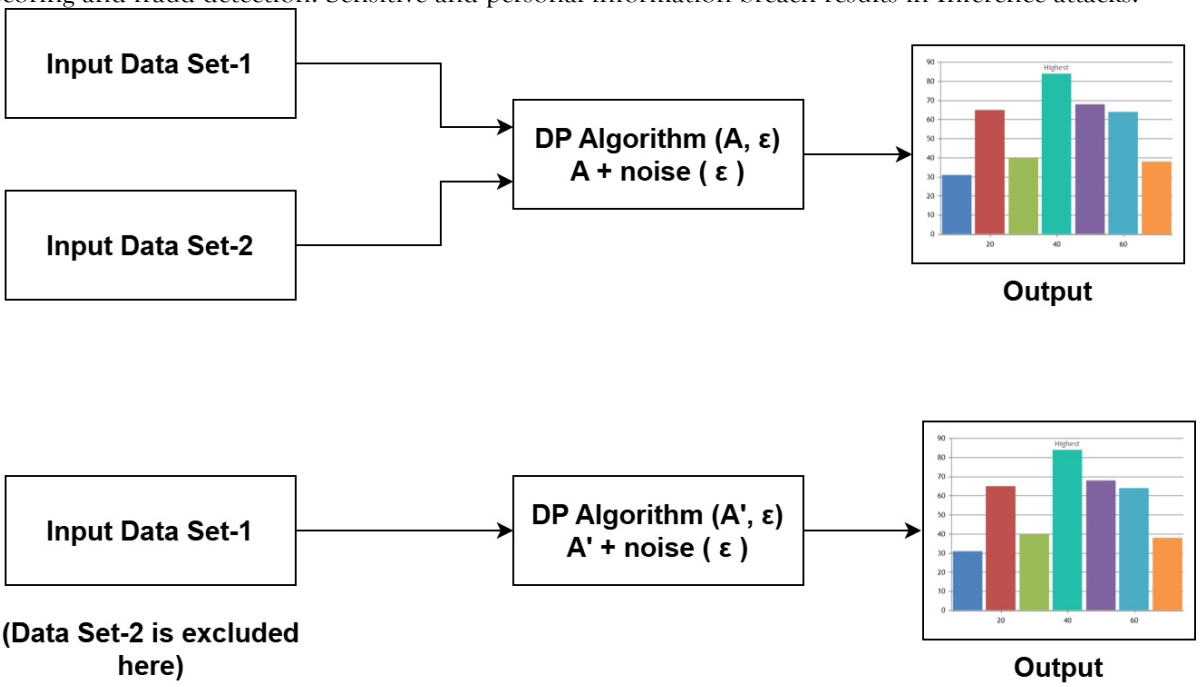

Figure :1 Mechanism of Differential Privacy technique

2.1.1 Differential Privacy

Differential Privacy is one of the widely used PPML methods to protect sensitive columns. Differential privacy is another PPML method that makes use of Laplace approximation to add noise. The method of adding noise with data is called data perturbation. The level of noise is proportional to the sensitivity, epsilon(**ε**) In the paper[2] , the author discusses about the challenges in balancing privacy and utility. **ε** is a privacy parameter where lower values of epsilon strengthens privacy but may decrease the utility. **Epsilon** is the key privacy parameter in Differential Privacy method and quantifies the privacy loss.

- **Lower ε** = stronger privacy, as the noise added to the results is larger.
- **Higher ε** = weaker privacy, as the noise added is smaller and the output is more accurate.

From Figure: It is apparent that the input dataset1 and input dataset2 is considered for analysis. A DP mechanism or algorithm is applied to the dataset1 (I1) and dataset2 (I2) along with the random noise (**ε**) which results in output (O1) (Equation: 1). another dataset consisting of only dataset1 (I1) is taken by excluding dataset2. DP mechanism is now applied on this along with the random noise (**ε**). The output generated is O2 (Equation:2). Both the outputs O1 and O2 will be probabilistically similar. The change in output depends on **ε**.

Equation 1: DPM(I1+I2+ ε) =O1

Equation 2: DPM(I1+ ε) =O2

Another parameter of Differential privacy is **Delta (δ) - The Probability of Privacy Breach** that indicates the probability value of violating the differential privacy.

In some mechanisms like **Gaussian Differential Privacy**, **σ** refers to the standard deviation of the noise that is added to the data. For **Gaussian noise**, the noise added to each query result is drawn from a Gaussian (normal) distribution with the average value of 0 and standard deviation **σ**. The noise helps to obscure the impact of any individual data point in the dataset. The larger the **σ**, the more noise is added, and the more privacy is guaranteed (but at the cost of utility).

In **Gaussian Differential Privacy**, **σ** typically depends on the sensitivity of the function being used (i.e., the impact of a single data point on the output of the function), and the chosen value of **ε** and **δ.**

Local Differential privacy ensures privacy at the input source level by assuring the data owner about the security and is well applicable in health care applications.

2.1.2 Homomorphic encryption

Homomorphic encryption provides security of data that is used in various applications like health care, financial sector, secure voting, fraud detection. It encrypts the input data used in machine learning models thus protecting sensitive information. It reduces unauthorized access and provides trust to the clients.

Homomorphic Encryption consists of the following phases:

Key derivation: The input data that could be private personal information needs security. Homomorphic encryption algorithm is applied o

n this data which produces an encrypted data. This encrypted data is secure and ensures input privacy. Machine learning model is applied on the encrypted data or the cipher text to produce a predicted output that is in encrypted form.(figure 2.a)

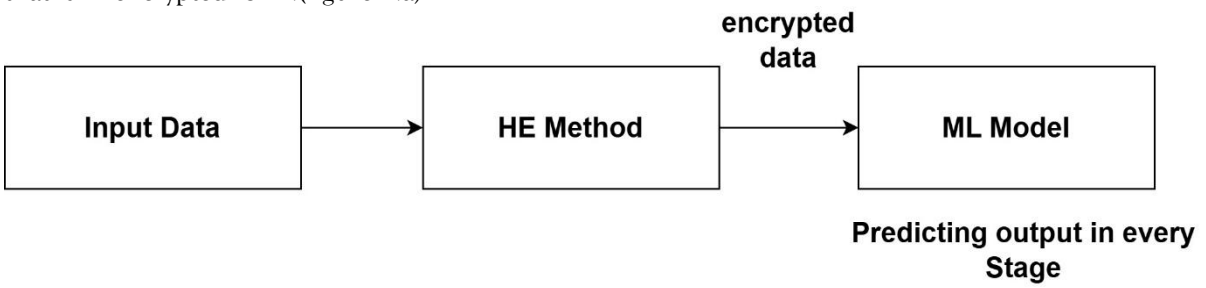

Figure:2.a Key Derivation and Encryption

Figure: 2.b Key Deciphering

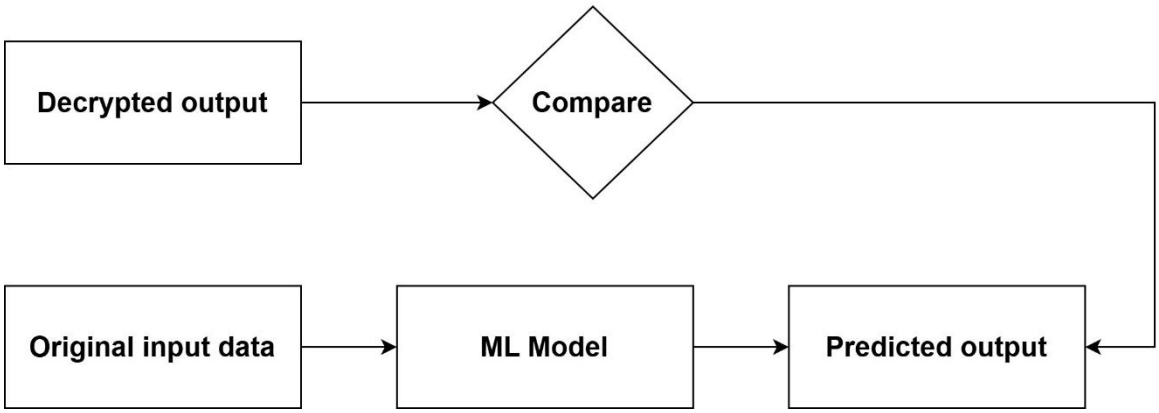

Figure: 2.c Comparison of Decrypted output and original input data.

Key deciphering: In this phase, Homomorphic decryption algorithm is applied on the Predicted result which is in encrypted form. The result is in decrypted form. (Figure 2.b).

Comparison: From figure 2.c, Machine Learning model is applied on the original data to produce an output, X. The decrypted output is compared with the output X to evaluate the accuracy of the results of Homomorphic encryption.

Homomorphic encryption provides security to the input data by encrypting it. Fully Homomorphic Encryption (FHE), Somewhat Homomorphic Encryption (SHE), and Partial Homomorphic Encryption (PHE) are the few techniques that allow computations on encrypted data. With reference to the paper [3], (Privacy-Preserving Chaotic Extreme Learning Machine with Fully Homomorphic Encryption), PHE provides only similar kinds of operations, either additions or multiplication can be performed any number of times on encrypted data. RSA, multiplicative homomorphism, Paillier (additive homomorphism) are examples of Partial Homomorphic Encryption.

Limited number of additive or multiplicative operations are performed in Somewhat Homomorphic encryption. Polycracker is an example for Somewhat Homomorphic encryption. Fully Homomorphic Encryption permits unlimited number of additive and multiplicative computations on the encrypted data. Gentry scheme which is lattice based is an example for FHE Microsoft's SEAL (simple Encrypted Arithmetic Library) is an implementation of CKKS (Cheon Kim Kim Song) that can approximate operations on encrypted real numbers. The challenges in HE is computational overhead and the CPU time for encrypted data is slow.

After a model has been trained, FHE is used to deduce inference (e.g., making predictions) on encrypted data, ensuring that the privacy of the data remains protected throughout the process.

In conclusion, FHE is a potential technology for enabling privacy-preserving computations across different disciplines, particularly where sensitive data needs to be processed securely. Its applications span healthcare, finance, cloud computing, and use cases that need strong privacy guarantees while still enabling useful computations on encrypted data. Homomorphic encryption provides security against poisoning attacks (Figure: 3) that ensures integrity of the Training dataset.

2.1.3 Secure Multiparty Computation (SMPC)

Privacy-Preserving Data Sharing among Multiple Parties: Multiple parties can use FHE to jointly perform computations on their private data (e.g., cross-enterprise data analysis) without exposing their individual datasets. Different sectors like banking, Insurance Companies, and Healthcare providers use FHE to analyse consumer data for fraud detection or market analysis without revealing their customers' private information.

Example: (Secure voting system). **Encrypted Ballot Counting**: In electronic voting systems, FHE serves to carry out tallying of votes while keeping the votes encrypted. Each vote can be encrypted before being cast, and the system can then perform homomorphic addition (i.e., counting the votes) on the encrypted data, ensuring voter privacy while allowing the correct final tally to be computed.[4] In the paper Secure-E-Voting System implementation using CryptDb, a method is proposed which does not disclose any information to the intruders at any level of polling system and hence the outcome of Voting process can be achieved using online system with Security, Confidentiality and integrity. With the help of CryptDB the system protects the information like the outcome of votes from the malicious administrator who tries to influence the voters in the voting process. This is a useful PPML technique to maintain security and privacy of user's data.

2.1.4 Federated Learning

Federated Learning facilitates collaborative training where data remains local on the devices and only model refinements are shared. This ensures that's confidential data adheres to not leave its storage thus maintaining privacy. In the paper [5] propose a secure Multi-Party Computation (secure aggregation) based Federated Learning framework. In a distributed learning model, the secure aggregation can protect the gradient information of each user's model. In the paper [6] the author summarized the major attacks like poisoning attacks, backdoor attacks, adversarial example attacks, model theft, and recovery of sensitive data. All the threats, defence techniques are analysed.

The Findings of the paper [7] include a classification of Privacy Preserving methods like SMPC, HE, DP, FL and garbled circuits, aggregation of the techniques to withstand the privacy threats. Also, the shortcomings of the approaches, further scope of challenges that arise to ensure privacy of users and data is presented in the said paper.

In the article, [8], a comparative study is made relevant to the threats, methods to solve the issues. Data anonymization is addressed by encryption techniques which results in reduction of leakage of sensitive data used for facial recognition systems. Differential privacy in AI Models is addressed by using various

differential privacy protocols that results in adequate protection of individual privacy during the training phase of AI models. In the paper [9] ethical analysis is made on issues like public safety and security which are the fundamental rights in liberal democracies like Australia, United States. Ethical principles are applied to bridge the gap between the methods applied to preserve privacy of data and democratic accountability on the other hand. These support the Human Data interaction principles (legibility, agency, negotiability) and the regulatory requirements like data protection, privacy, transparency and data minimization.

2.2 Explainable AI

Explainable AI relates to the task of explaining the output of a Machine Learning model to the stakeholders. Explainable AI provides insight in two aspects. One, Explainable AI enables the naive users understand the decisions of the models and two, the impact of the variables used to predict the outcome. For instance, in the prediction of covid-19 disease, several factors like fever, cold, cough, were taken into consideration. Explainable AI helps in finding out which symptom had a major impact on the occurrence of the disease. In Image Classification, Explainable AI can explain why an image was defined as a cat, emphasizing the specific pixels that contributed to the decision. The need of Explainable AI arises from the wide usage of accurate and complicated AI models. Such models make it difficult for the end users to understand and interpret the decision making process. The four principles of Explainable AI are elucidated in the Figure 2.d:

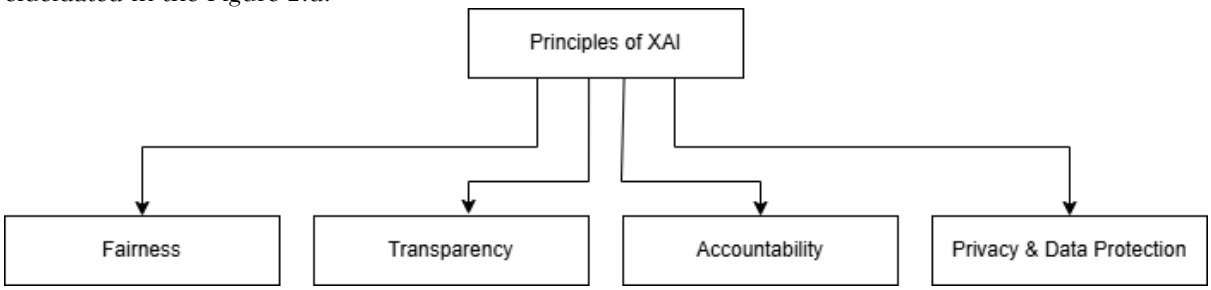

Figure 2.d: Principles of XAI

The principles of XAI achieve model interpretability, feature attribution and model explainability.

The two approaches to implement Interpretable AI are:

Rule-based approach: Explainable AI uses Rule based algorithms like decision trees, fuzzy-rule based systems to interpret the decisions of the models.

Post-hoc approach: in this approach Explainable AI interprets the decision making of black-box models by adding surrogate models. The two types of post-hoc approach are:

Local Explanations: The focus here is to explain the reason for a single prediction for a particular feature or an attribute.

Global Explanations: The focus here is to comprehend the process of the model learning from the data. The process describes the overall behaviour of the complex back-box models that use high dimensional space.

Model agnostic tools like Local Interpretable Model Agnostic Explanation (LIME) and Shapley Additive Explanations (SHAP) are used to explain the decision making of the models.

Model-agnostic techniques in Machine learning refer to the methods that can be applied to any machine learning model, irrespective of its architecture or type. These techniques are typically independent of the model's specific structure, making them flexible for various tasks. A model-agnostic technique aims to improve or modify how a model operates, without needing to make changes to the underlying model itself.

LIME is local as it explains individual predictions using any simple model. For example, it can predict whether a customer will purchase a product based on his frequency of his visit to the website (local explanation). SHAP provides both global (overall features significance) explanation and local (individual feature importance) explanations. Shapley values from game theory are used to identify the importance of each feature in prediction.

In the paper [10], the importance of explainable AI is suggested. For the applications that use complex deep learning approaches and black box models, the mechanism used to explain the results is not human interpretable. Also, it is difficult to understand the cause of an unexpected action taken by the black box model. In such situations, we need a more Transparent, interpretable AI systems. In the paper [11], the author focusses on the significance of Explainable AI to enhance the performance of the output of

Machine Learning models. The paper discusses about a framework that uses XAI to improve the decision-making capabilities of pre-trained models without building the model again from its inception. Improving the decision capabilities with the help of XAI reduces exploratory attacks. . In the paper [12], the author discussed various challenges of explainable AI in deep learning and interdisciplinary research directions.. Model agnostic methods like Local Interpretable Model-Agnostic Explanations (LIME) and Shapley Additive Explanations (SHAP) are discussed that explain the importance of input features that effect the decision of the ML models. The paper contributed in explaining the use of XAI methods in various applications and also its social impact.

The major focus of the paper [13] is on four aspects namely data explainability, model explainability, post hoc explainability and assessment of explanations that refine the model performance in terms of robustness and also ensure trust among the stakeholders.

The goal of the paper [14] is to provide an exhaustive explanation ontology which considers different forms of data like tabular data, images, text and time series data. The paper addresses issues like transparency by using intrinsically explainable design methods that return the reason for decision making of the model. Global methods are used to explain the logic of the black box model.

The aim of the paper [15] was to provide human interpretable explanations to the non - expert users. The aim was achieved after conducting a series of interviews on producing industries and related them with the academic XAI research.

The paper [16] presents a taxonomy of interpretability methods for black-box models, white- box models, that ensure transparency and lastly about analysing sensitivity of the model predictions. LIME and SHAP approaches explain the significance of features in predictions of the output and also these methods work on almost any type of data.

3. Motivation for Integration of XAI and PPML

The need for transparency, human interpretability and trustworthiness is the motivation for analysing the importance of integrating both Explainable AI and Privacy Preserving Machine Learning techniques.

For example, a Machine Learning Model like Support Vector machine (SVM) used for text classification may label an email as spam or not spam based on hundreds of features (word, metadata) but the decision-making process is difficult to understand without tools like feature importance analysis. The black-box modes that uses many parameters, decision paths interact in a non-linear way thus making it difficult to track how individual inputs influence the predicted output. Factors like lack of transparency and high dimensional- feature space make it difficult for end users to understand and trust the results of the models.

3.1Privacy Preserving XAI in Federated Learning

Federated Learning facilitates collaborative training where data remains local on devices and only model updates are shared. This ensures that confidential information does not leave its local storage thus maintaining privacy.

After the model training in a federated set up, techniques like LIME/SHAP generate explanations based on the refinements of the model. Federated environment could leverage encrypted model updates and distributed XAI techniques, allowing each party to compute an explanation for its local model without exposing the data.

3.2 DP with XAI

In a loan approval system. Differential Privacy approach ensures that no single applicant's data can be identified but the system still explains, in an aggregated way, the denial of a particular loan. The explanation relates to the general features (ex: credit score, income level) that contributed to the decision. XAI techniques can explain how the model ensures differential privacy such as showing how noise is injected into the predictions. XAI focuses on certain transformations (like noise addition) that has been applied to the predictions to guarantee privacy.

3.3 XAI with Trusted Execution Environments

TEE are Hardware based isolated environments (ex: Intel SGX, ARM trust zone) where computations can be performed securely. TEE's are leveraged to a XAI algorithm on encrypted data within a secure enclave, raw data and explanations of the outcome are protected.

LIME/SHAP techniques within a TEE are used to securely generate interpretations of the model predictions while ensuring that confidential information data never leaves the enclave. The model itself can be aggregated, explanations can be computed without exposing either the data/explanation.

3.4 Model -Agnostic explanation using secure Aggregation

Approaches like LIME/SHAP explain the predictions of any Machine Learning model, irrespective of its architecture. These techniques can still be applicable if secure aggregation methods are used. The secure aggregation methods provide individual contributions to the explanation (ex: impact of a feature on a prediction) can be aggregated securely without revealing private information. The contributions from different participants or sources of data can be integrated in an encrypted format resulting in an aggregated explanation.

4. Challenges in PPML

a) Privacy-utility Trade-off: To develop a model that maintains a balance between privacy guarantees (Differential Privacy) and the utility (accuracy) of a model.

b) Scalability and Efficiency: To evaluate the efficiency of the model in real-world applications, considering the overhead of encryption and secure computations.

c) Computational Overhead: Homomorphic encryption leads to significant overhead in both training and inference time. To ensure optimization of training and inference time for practical use is a challenge.

d) Security vs. Performance Trade-offs: To build a model that strikes a balance between the level of encryption and maintaining model accuracy and interpretability.

e) Model Adaptability: To ensure that PPML techniques do not hinder the adaptation or fine tuning of models over time.

5. Challenges in XAI

a) Lack of standardization: The absence of universal metrics to measure explainability is a direction to research. Different users might interpret same explanation differently. To develop a standard approach that meets the needs of diverse users is a challenge. To identify universal metrics that measures explainability contributes to standardization.

b) Complexity of Models vs Interpretability trade-off: Numerous deep Learning models implement complex procedures like high dimensional space, non-linear methods that act as black-box models. Making these models explainable without compromising performance of the model is a challenge.

c) Scalability: Explainable solutions that work for small data sets do not necessarily produce same outcomes for a high dimensional space. Providing explainable solutions for large scale models is still an open challenge.

d) Trustworthiness: The usefulness of the Explanations or the approaches used in understanding the black-box models provides trust to the stakeholders. Delivering trustworthy explanations to the users is a research concern.

e) Human Data Interaction Ethics and Principles: Legibility, agency, negotiability are the principles that require attention. The task is to explain the technology to the public and the implications of their choice while preserving privacy.

6. Challenges in Privacy -Preserving ML and Interpretable Machine Learning

a) The fundamental issue is to ensure that both privacy (through secure computations on encrypted data) and transparency (through explainable AI techniques) are preserved simultaneously.

b) To implement secure aggregation Techniques for model-agnostic XAI methods ensuring that Interpretability and privacy are maintained simultaneously is a research challenge.

c) Developing privacy-preserving versions of XAI Techniques that can integrate Differential privacy method to provide secure, accurate and interpretable solutions is still an open research problem.

d) To develop models using XAI Techniques like LIME/SHAP to work with the ML models that operate on encrypted data and produce secure predictions is an open research issue.

e) The tolerances between privacy, model performance and explainability when integrating these approaches in practical applications is a challenge.

f) The impact of surrogate frameworks on the interpretability of complex models when operating on secure data is a research problem that needs attention.

7. CONCLUSION

The need for data security and trustworthy AI grows as the technology is used widely across various domains like health care, financial sector, cybersecurity, and recommendation systems that involve automation and decision making. This paper covered various data threats, Privacy Preserving Machine

Learning techniques, and Explainable AI approaches in the background and foundation section. The study discusses the Homomorphic Encryption technique, Differential privacy and Simple multi-party computation techniques. In addition to the need of privacy preserving, the major aspects discussed are Interpretability, Trustworthiness, and Transparency that can be achieved through Explainable AI.

Three major concerns are discussed in this comprehensive survey. Firstly, the various PPML approaches and challenges to ensure data security are elicited. Second the XAI techniques that explains the significance of transparency, communication of meaning from data is explained. Third, the major challenges and future directions of integration of PPML and XAI are summarized.

This survey paper gives open directions and challenges to the researchers who want to provide trustworthy ML and AI models without compromising the trade-off between privacy and utility of Machine Learning models.

REFERENCES

[1] Ensuring data privacy using Machine Learning for Responsible data science. Jena, M.D., Singhar, S.S., Mohanta, B.K., Ramasubbareddy, S. (2021). Ensuring Data Privacy Using Machine Learning for Responsible Data Science. In: Satapathy, S., Zhang, YD., Bhateja, V., Majhi, R. (eds) Intelligent Data Engineering and Analytics. Advances in Intelligent Systems and Computing, vol 1177. Springer, Singapore.

[2] Privacy-Preserving Machine Learning Techniques, Challenges And Research Directions Deval Parikh1, Sarangkumar Radadia2, Raghavendra Kamarthi Eranna3

[3] Privacy-Preserving Chaotic Extreme Learning Machine with Fully Homomorphic Encryption Syed Imtiaz Ahamed1,2 and Vadlamani Ravi* 1Centre for AI and ML, Institute for Development and Research in Banking Technology Castle Hills, Masab Tank, Hyderabad 500057, India 2 School of Computer and Information Sciences (SCIS), University of Hyderabad,

[4] Secure e-voting using CryptDB

[5] Practical Secure Aggregation for Federated Learning on User-Held Data. Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, Karn Seth

[6] Machine Learning Security: Threats, Countermeasures, and Evaluations M. Xue, C. Yuan, H. Wu, Y. Zhang and W. Liu, "Machine Learning Security: Threats, Countermeasures, and Evaluations," in IEEE Access, vol. 8, pp. 74720-74742, 2020, doi: 10.1109/ACCESS.2020.2987435.

[7] Privacy-preserving artificial intelligence in healthcare: Techniques and applications. Author links open overlay panelNazish Khalid [a], Adnan Qayyum [a], Muhammad Bilal [b], Ala Al-Fuqaha [c], Junaid Qadir [d]

[8] Toward a Comprehensive Framework for Ensuring Security and Privacy in Artificial Intelligence. . William Villegas-Ch * and Joselin García-Ortiz

[9] Ethical application of biometric facial recognition of Technology Marcus Smith1 · Seumas Miller Received: 6 October 2020 / Accepted: 23 March 2021 / Published online:

[10] 2017 the 2nd IEEE International Conference on Cloud Computing and Big Data Analysis An Exploration on Artificial Intelligence Application: From Security, Privacy and Ethic Perspective

[11] Towards a general framework for improving the performance of classifiers using XAI methods arXiv:2403.10373v1 [cs.LG] 15 Mar 2024 Andrea Apicella, Salvatore Giugliano, Francesco Isgr` o, Roberto Prevete Department of Electrical Engineering and Information Technology, University of Naples Federico II

[12] Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. Author links open overlay panelLuca Longo [1 2], Mario Brcic [3], Federico Cabitza [4 5], Jaesik Choi [6 7], Roberto Confalonieri [8], Javier Del Ser [9 10 11], Riccardo Guidotti [12], Yoichi Hayashi [13], Francisco Herrera [11],
Andreas Holzinger [14], Richard Jiang [15], Hassan Khosravi [16], Freddy Lecue [17], Gianclaudio Malgieri [18], Andrés Páez [19 20], Wojciech Samek [21 22 23], Johannes Schneider [24], Timo Speith [25 26], Simone Stumpf [27]

[13] Explainable Artificial Intelligence (XAI): What we know andwhatisleftto attain Trustworthy Artificial Intelligence Sajid Alia, Tamer Abuhmedb,*, Shaker El-Sappaghb,c,d, Khan Muhammade,*, Jose M. Alonso-Moralf, Roberto Confalonierig, Riccardo Guidottih, Javier Del Seri,j, Natalia Díaz-Rodríguezk, Francisco Herrerak

[14] Benchmarking and survey of explanation methods for black box models Francesco Bodria1 · Fosca Giannotti1 · Riccardo Guidotti3 · Francesca Naretto1 · Dino Pedreschi3 · Salvatore Rinzivillo2

[15] Decker, T., Gross, R., Koebler, A., Lebacher, M., Schnitzer, R., Weber, S.H. (2023). The Thousand Faces of Explainable AI Along the Machine Learning Life Cycle: Industrial Reality and Current State of Research. In: Degen, H., Ntoa, S. (eds) Artificial Intelligence in HCI. HCII 2023. Lecture Notes in Computer Science(), vol 14050. Springer, Cham. (not in sci-hub)

[16] Explainable AI: A Review of Machine Learning Interpretability Methods Pantelis Linardatos ∗ , Vasilis Papastefanopoulos and Sotiris Kotsiantis