Journal of
Cybersecurity and Privacy

MDPI

Article

A Comparative Analysis of Defense Mechanisms Against Model Inversion Attacks on Tabular Data

Neethu Vijayan *, Raj Gururajan and Ka Ching Chan *

School of Business, University of Southern Queensland, Queensland, QLD 4350, Australia; raj.gururajan@unisq.edu.au
* Correspondence: neethu.vijayan@unisq.edu.au (N.V.); kc.chan@unisq.edu.au (K.C.C.)

Abstract

As more machine learning models are used in sensitive fields like healthcare, finance, and smart infrastructure, protecting structured tabular data from privacy attacks is a key research challenge. Although several privacy-preserving methods have been proposed for tabular data, a comprehensive comparison of their performance and trade-offs has yet to be conducted. We introduce and empirically assess a combined defense system that integrates differential privacy, federated learning, adaptive noise injection, hybrid cryptographic encryption, and ensemble-based obfuscation. The given strategies are analyzed on the benchmark tabular datasets (ADULT, GSS, FTE), showing that the suggested methods can mitigate up to 50 percent of model inversion attacks in relation to baseline models without decreasing the model utility (F1 scores are higher than 0.85). Moreover, on these datasets, our results match or exceed the latest state-of-the-art (SOTA) in terms of privacy. We also transform each defense into essential data privacy laws worldwide (GDPR and HIPAA), suggesting the best applicable guidelines for the ethical and regulation-sensitive deployment of privacy-preserving machine learning models in sensitive spaces.

Keywords: differential privacy; federated learning; model inversion attack; privacy-preserving machine learning; tabular data security

check for updates

Academic Editor: Martin Gilje Jaatun

Received: 9 June 2025
Revised: 29 July 2025
Accepted: 16 September 2025
Published: 2 October 2025

Citation: Vijayan, N.; Gururajan, R.; Chan, K.C. A Comparative Analysis of Defense Mechanisms Against Model Inversion Attacks on Tabular Data. *J. Cybersecur. Priv.* **2025**, *5*, 80. https://doi.org/10.3390/jcp5040080

1. Introduction

The application of machine learning (ML) across various industries has transformed operations and enables faster, more accurate, and predictive decision-making. In medicine, machine learning enables progress in personalized treatment by helping to customize plans based on genetic profiles [1–3]. Recommendation engines based on ML are used on e-commerce sites to forecast consumer behavior and improve their quality [4,5], whereas in the financial sector, ML is applied to algorithmic trading, fraud detection, and risk analysis [6]. Outside these areas, ML is also being used in smart infrastructure and industrial automation, where it is used to schedule and optimize production, frequently involving the creation and utilization of synthetic data to augment sparse raw data. Discrete event simulation (DES) and generative modeling techniques allow for ML systems to be trained on artificial information that is statistically similar to the real operational data and obviate the need to expose sensitive information directly [7,8].

As the use of ML systems grows in data-sensitive areas, privacy concerns have become more urgent. Unregulated handling of personal, proprietary, or confidential data during model training makes systems prone to accidental data leaks. Among recent threats, model inversion attacks are especially critical: even without direct access, the output from trained models can reveal private information about the training data. Remarkably, this was demonstrated by Fredrikson et al. [9], who reverse-engineered a pharmacokinetic ML model to gain partial access to a patient's genomic profile, highlighting the dire insecurity of such models, which unintentionally reveal the data they are supposed to safeguard.

Amid these risks, one studied solution is generating synthetic data as a privacy-protecting measure, especially for tabular data. This involves statistically replacing or augmenting real records with synthetically similar samples to lower the risk of disclosure without losing utility. However, producing data alone is not enough: generative models can unintentionally memorize or mimic real people's patterns, making it impossible to prevent inference attacks entirely. Therefore, strong, layered privacy-protection methods like differential privacy, federated learning, and advanced cryptographic barrier nets are essential, particularly when safeguarding sensitive data, as the international legal framework requires protective measures against such data.

To give an example, in healthcare, synthetic data functions to remove personally identifiable information (PII) and synthesize datasets that are not directly associated with individuals [10]. In production, artificial data secures competitive aspects and trade secrets. Still, the two areas demand more defense mechanisms than synthetic data to prevent skilled adversarial inference, as creating and evaluating advanced privacy-preserving ML techniques is of value. In this regard, the protection of ML systems based on structured/tabular data containing sensitive personal, financial, or operational information is critically important. Since model inversion attacks pose a potential threat, effective defense mechanisms should not only prevent these attacks but also comply with existing data protection laws (including GDPR and HIPAA), making them suitable for use in high-stakes applications. Although extensive research has been conducted, the majority of the work only assesses either specific defenses or targets unstructured data, resulting in a lack of comparative, overall assessments in tabular contexts.

1.1. Goals and Scope

This paper proposes and systematically analyses a set of privacy-protecting methods of ML models in tabular settings. Our comparative study unites and compares several defense mechanism solutions, such as differential privacy, federated learning, hybrid encryption, adaptive noise injection, and ensemble obfuscation, on an equal experimental basis. To evaluate the potential privacy leakage suppression of each strategy (measured by inversion accuracy, mutual information reduction, and reconstruction error) and the preservation of model utility (accuracy, F1 score, and latency), we utilize real-world tabular datasets in the areas of healthcare, finance, and social science. This comparative analysis exposes trade-offs, domain-restricted boundaries, and pragmatic suggestions on how to advance the secure and ethical use of ML in sensitive aspects of data.

To enhance the visualization of the privacy risks associated with deploying machine learning models over tabular data, Figure 1 presents a comparison of a typical ML workflow with an adversarial one, where a model inversion attack is employed using tabular data. In the traditional case, tabular data would be gathered, a model would be trained and established, and users would submit a query to obtain predictions. As a contrast, the adversarial setup illustrates that an attacker can obtain query access to the deployed model, feed it with inputs, and receive outputs to reverse-engineer sensitive attributes on the training data, leading to a major privacy violation [11].

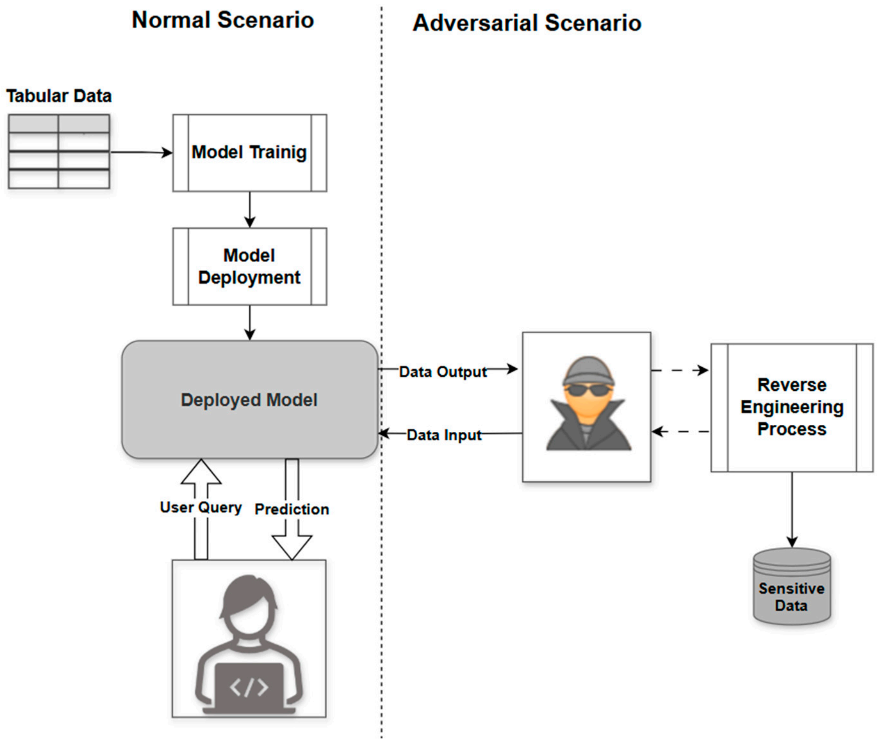

Figure 1. Comparison between a standard ML workflow and a model inversion attack scenario.

1.2. Contributions

The primary contributions of this work are:

- Comparative analysis of defense mechanisms: We evaluated various defensive techniques—differential privacy, federated learning, adaptive noise injection, hybrid encryption, and ensemble obfuscation—using tabular datasets to assess their effectiveness.
- Comparison across different datasets: We conducted an in-depth analysis across multiple datasets to understand how these defense techniques affect different conditions and data types.
- Data compliance: We interpret each defense within the context of major data privacy laws worldwide, such as GDPR and HIPAA, and recommend the most appropriate guidelines to ensure ethical and regulation-compliant deployment of privacy-preserving machine learning models in sensitive sectors.

1.3. Paper Structure

Section 2 reviews related work and identifies a gap analysis that motivates our study. Section 3 details datasets, threat model, defenses, metrics, and the experimental protocol. Section 4 presents results including baseline vulnerability, privacy–utility curves, comparative defense performance, and domain generalization. Section 5 discusses limitations and future directions; Section 6 offers the conclusion.

2. Literature Review

The growing use of machine learning models in sensitive industries like healthcare, finance, and infrastructure has increased the need for security against privacy threats. One of the most challenging attacks is the model inversion attack, where adversaries are only able to use outputs of a model to infer sensitive training data, even when raw data is never directly revealed [12]. The threat level is extremely high when it comes to structured tabular data with highly identifiable attributes. This sensitive situation is not only related to what the model generates but also to the data structure, open API endpoints, and the complexity of how features interact in the real world. Therefore, becoming familiar with the methods used by adversaries and the mitigation techniques is essential for securing ML in these environments.

2.1. Foundational Inversion Attacks

An early work by Fredrikson et al. [9] showed that securely trained models could still leak sensitive information (e.g., genomic information) when the attacker knew some information about the training inputs and could access the model outputs. Amini Gougeh [13] and Kulkarni & Bhambani [14] followed up with this and considered adversarial image perturbations (FGSM, PGD) to tabular and medical cases, as they demonstrated that such manipulations could expose very sensitive information in patient or financial data.

The primary cause of a high risk of privacy in tabular data has to do with its semantics. Such features as income, medical history, or age can be directly interpreted in the sense that attacks can be easier to pursue than on the abstracted pixel values of images. This semantics transparency further makes it difficult to use inter-feature correlations, which the inversion attacks depend on. Accordingly, tabular adversarial manipulation may be discrete and therefore hidden but is frequently considerably drastic.

To counter these threats, various defense mechanisms have been devised with the scope of bringing privacy without compromising security. Tables 1 and 2 break down these mechanisms into differential privacy, federated learning, hybrid encryption, adaptive noise injection, and ensemble obfuscation in terms of their core purpose, implementation phase, and prevention impact. More importantly, Table 1 is here combined with a mapping of the most relevant prior works, including both an at-a-glance overview and a contextual linkage of each strategy and literature contribution.

Table 1. Summary of Model Inversion Attacks Research and Their Limitation.

Study	Data Type	Focus	Application Domain	How It Mitigates Inversion Attacks	Limitation or Gap
Fredrikson et al. [9]	Tabular	Attack	Healthcare	Model inversion via confidence scores	Requires partial input knowledge; early-stage framework
Amini Gougeh [13]	Image	Attack	Medical Imaging	FGSM, PGD adversarial attacks	Focused on CNNs; limited translatability to tabular structures
Kulkarni & Bhambani [14]	Tabular	Attack	Finance, Health	Regional adversarial manipulation	Demonstrates risk; lacks defense validation
Brendel & Bethge [15]	Tabular/Image	Attack	General	Optimization-based perturbation	Effective across domains; lacks targeted mitigation

As summarized below, most of the earliest model inversion literature was concerned with an unstructured domain. However, interpretable features and regulations make tabular datasets, which dominate in healthcare, financial, and administrative systems, deal with even greater privacy threats directly. Most importantly, the majority of defenses devised for images or text are not easily transferable to tabular data or lead to significant reductions in model utility.

The same research has noted that models trained with synthetic or semi-synthetic data would remain vulnerable when the generative process uses patterns in the original data [16]. Synthetic data should be audited carefully in highly structured environments to avoid leakage by mistake. Stronger defense techniques, including differential privacy (DP) [17] that add noise into training, but at high noise levels they can hurt model performance.

A more recent promising solution that is currently under consideration is Context-aware privacy budgets, where the noise is allocated based on the sensitivity of the specific features.

Table 2. Summary of Defense Mechanisms Against Model Inversion Attack.

Study (Author, Year)	Data Type	Focus	Application Domain	How It Mitigates Inversion Attacks	Limitation or Gap
Zhang et al. [18]	Tabular/Image	Defense	Medical, Finance	Obfuscates feature-target relationships; limits attribute leakage	Uniform noise may reduce utility significantly
Li et al. [19]	Tabular/Image	Defense	Federated Medical Models	Prevents raw data sharing; reduces central exposure	High overhead: context-sensitive calibration needed
Alzubaidi et al. [20]	Image	Defense	Medical Imaging	Uses multiple models for prediction; dilutes direct feature influence	Not directly usable for structured/tabular ML

Also, Li et al. [19] also addressed the privacy issue of deep learning related to military and medical images, where even small perturbations of the input can expose important hidden variables. As Guo et al. [21] showed, causing an oversized impact on the outputs by manipulating a single feature (such as in the case of One-Pixel Attack), which has direct implications in the context of tabular ML because a single feature like smoker, married, etc., can lead to major changes in prediction.

Regardless of the numerous individual developments, the literature indicates that a multi-faceted defense for tabular ML systems is necessary. There should be no post hoc anonymization; instead, proactive measures such as model diversification, synthetic data verification, different privacy calibration, and online query auditing are required. Interpretability is the key feature that makes tabular ML valuable, as well as vulnerable: an apparent decision process makes targeting sensitive attributes more accessible.

2.2. Gap Analysis

Although recent developments have concentrated on specific aspects of model inversion and privacy protection, several gaps still exist. Most comparative studies focus on unstructured data or consider only one defense method. The unique features of tabular ML, such as interpretability of features, domain control, or the need to balance utility and privacy, are rarely addressed comprehensively in a single empirical study. Additionally, state-of-the-art (SOTA) techniques, such as local differential privacy, transformer-based attacks on inversion, and adaptive defense tuning, are underrepresented in tabular benchmarks. Terms like operational regulatory compliance (e.g., GDPR, HIPAA) are discussed only in general terms, without technical validation.

The final concern is that standard and effective methods for measuring risk have not been developed within the context of model inversion attacks. Most current discussions of privacy threats are descriptive, so the exposure and loss of privacy cannot be effectively quantified or compared, and neither can the effectiveness of gender-specific protective measures be measured quantitatively. The work directly addresses these restrictions by benchmarking five complementary defenses on both synthetic and real tabular datasets, aligning with technical and regulatory requirements, and providing concrete guidance on privacy-utility trade-offs in high-stakes ML deployments.

2.3. Recent Developments in Data Privacy and Inversion Attacks

Recent research has broadened the scope of privacy-preserving machine learning for tabular data. For instance, local differential privacy (LDP) techniques tailored for structured datasets have demonstrated practical compromises between utility and privacy in decentralized settings [22,23]. Furthermore, approaches that dynamically adjust noise or obfuscation levels during inference are gaining traction as practical strategies to counteract evolving adversarial threats [24].

While this study centers on tabular machine learning models, it is essential to recognize that similar privacy issues also exist with large language models (LLMs) [25,26]. Recent studies have shown that LLMs can be susceptible to training data leaks through prompt-based inversion or membership inference attacks [27–29]. Researchers have found that exposing sensitive data during fine-tuning poses risks through attacks such as membership inference, data extraction, and backdoor attacks [30]. They also examined defense methods such as differential privacy and federated learning during the fine-tuning process.

Additionally, Petrov et al. [31] demonstrated a gradient inversion attack called DAGER (Discreteness-Based Attack on Gradients for Exact Recovery), which can reconstruct entire text inputs from shared model gradients in federated LLM training. Lukas et al. [32] assessed attacks on GPT-2 models and defined three types of PII leakage using only API access to a language model. These findings emphasize that simple anonymization or basic differential privacy by themselves are inadequate against persistent LLM adversaries.

Despite differences in structure and data types, these models face common threats such as unintentional memorization and the risk of sensitive information being reconstructed. Future research could investigate whether multi-modal or hybrid defense strategies might provide valuable insights for both structured tabular models and unstructured language models.

3. Research Methodology

To develop a robust defense framework against model inversion attacks in tabular ML systems, this study adopts a structured, multi-phase experimental methodology. Each phase builds upon the previous one, enabling systematic evaluation of defense strategies in terms of both privacy protection and model utility. The process also includes regulatory alignment with GDPR and HIPAA standards.

The complete workflow is illustrated in Figure 2, and the methodology is detailed in the following subsections.

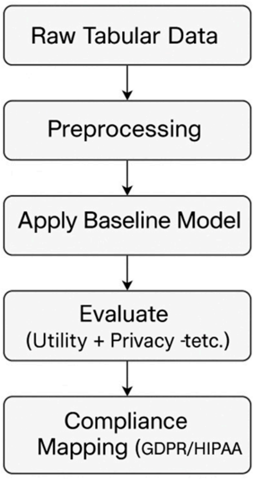

Figure 2. Methodology Workflow.

3.1. Attack Model

In our attack model, the adversary aims to reconstruct sensitive attributes using the outputs of a deployed model. Our main focus is on black-box inversion, where the attacker only has query access to the target model. The attacker trains a neural network inversion model that learns to predict the sensitive attribute by using the target model's outputs as inputs. The architecture employed is a multi-layer perceptron (MLP) with ReLU activations. We used class-weighted loss functions to address the imbalanced distribution of sensitive attributes. Performance metrics, such as accuracy, precision, recall, and F1 score, are calculated and averaged over five independent runs to report the mean and standard deviation. We primarily focus on a black-box threat model because it reflects realistic deployment scenarios in commercial and medical settings, where model internals are typically inaccessible.

3.2. Baseline Model Training

The initial step will be training baseline ML models (Decision Tree, Random Forest, Multi-Layer Perceptron) on three real-life tabular data sets: ADULT (UCI Income), GSS (General Social Survey), and FTE (Five Thirty-Eight). A summary of these datasets is provided in Table 3. They were chosen due to their distinctiveness and frequent analysis; the three categories are census, social science, and financial transaction data, which are considered baseline examples in privacy research and literature relative to ML.

Table 3. Dataset Overview.

Dataset	Features	Sensitive Attributes	Preprocessing Applied
ADULT	14	Income	Normalization, Encoding
GSS	12	Religious Affiliation	Scaling, Missing Value Handling
FTE	16	Transaction Purpose	Normalization, Feature Engineering

Despite the availability of several tabular datasets, these three provide a convenient combination of data scale, diversity, and existing privacy-sensitive attributes (income, religious affiliation, transaction purpose), which allows comparing the results with previous studies and state-of-the-art techniques. The missing values were imputed, and features were normalized using standard methods for categorical variables (e.g., one-hot encoding, z-score normalization). At this stage, no defense mechanisms were applied, and it was possible to measure the baseline vulnerability to attribute leakage and identify reference points for further defensive actions.

3.3. Defense Strategy Integration

At this second step, the different defense mechanisms are inserted and tested individually using the same models and datasets, allowing for direct comparison in an equalized fashion. The defense response measures are:

Differential Privacy (DP): The DP-SGD optimizer adds noise to the model gradients during training. The privacy budget epsilon is varied systematically (epsilon in {0.1, 0.5, 1.0, 5.0}) to examine the trade-off between privacy and utility, as found in other studies [24]. The outcome of model accuracy and the success of an inversion attack is noted at each value.

Equation for DP-SGD Noise Injection:

$$\theta_{t+1} = \theta_t - \eta\left(\nabla L(\theta_t) + N\left(0, \sigma^2 I\right)\right)$$

Here, η is the learning rate, L is the loss, and $N\left(0, \sigma^2 I\right)$ is the injected noise, which is a Gaussian.

Federated Learning: Mimics is a distributed environment with five client nodes, where Federated Averaging (FedAvg) is used over 50 communication steps [33]. The local data remains on each client's device; only the transmitted and aggregated gradients are shared, which accurately portrays privacy-preserving deployment environments in a realistic setting.

Hybrid Encryption: Elliptic Curve Cryptography (ECC) is used in combination with AES-128 to encrypt data both in transit and at rest symmetrically [34–37]. Both sensitive features and model parameters are encrypted prior to being stored or transferred. The computation overhead and latency are recorded.

Adaptive Noise Injection: A type of noise used in sensitivity scores for features and small features at risk, identified through mutual information analysis [38–40]. The approach allows for high-resolution adjustment of the noise impact while maintaining the overall usefulness of the model.

Ensemble Obfuscation: This combines (aggregates) predictions from different models (Random Forest, MLP, k-Nearest Neighbors) using a voting classifier [41]. This approach aims to obscure deterministic feature-to-output mappings and reduce overspecialization on sensitive patterns in a single model.

All defenses are built on open-source ML frameworks such as scikit-learn, TensorFlow Federated, and PyCryptodome, with every hyperparameter and setting documented for reproducibility. Clear records of dataset and model configurations are maintained for each table. Table 4 summarizes the configurations for baseline models and each defensive strategy, providing a concise overview of parameters, datasets, and specific implementation details.

Table 4. Model and Defense Configuration.

Technique/Model	Parameters	Dataset (s)	Notes on Configuration
Decision Tree	Gini, max depth = 10	ADULT, GSS, FTE	Baseline model for tabular classification
Random Forest	100 estimators, max features = auto	ADULT, GSS, FTE	Improved generalization and baseline ensemble comparison
MLP	3 layers, 128-64-32 nodes, ReLU, dropout = 0.2	ADULT, GSS	Used for benchmarking deep learning sensitivity
Differential Privacy	$\varepsilon \in \{0.1, 0.5, 1.0, 5.0\}$, noise on gradients	ADULT, FTE	Implemented using DP-SGD optimizer
Federated Learning	5 simulated clients, FedAvg, 50 rounds	ADULT, GSS	Used TensorFlow Federated; no raw data sharing
Hybrid Encryption	ECC key exchange, AES-128 encryption	ADULT, FTE	Encrypts at rest and in transit; additional computation overhead
Adaptive Noise Injection	Gaussian noise scaled by feature sensitivity scores	ADULT, GSS, FTE	Higher noise on sensitive features, low impact on performance
Ensemble Obfuscation	Voting classifier (RF, MLP, KNN)	GSS, FTE	Blurs attribution by averaging over diverse models

The hyperparameters used in the experiments were carefully chosen based on standard results from machine learning and privacy defense methodologies for tabular tasks. Specifically, the sizes of Decision Trees and Random Forests are widely referenced as benchmarks when working with structured data, similar to how Multi-Layer Perceptrons (MLPs) are more associated with deep learning in tabular data setups. Additionally, the limits on parameters for defense mechanisms, including privacy budgets (which is 0), the magnitude of adaptive noise injection, and the number of clients in federated learning, are based on values commonly disclosed and practically achievable, as described in the existing literature. This selection of parameters aims to ensure that the solution achieves a good balance between privacy protection effectiveness and its deployment feasibility in real-world applications.

3.4. Evaluation and Regulatory Mapping

Both utility and privacy risk metrics are used to assess the models and defenses. For utility, common measures include accuracy, precision, recall, and F1 score.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP} + FN$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

For privacy risk, metrics include Inversion Accuracy, which is the success rate at which adversaries reconstruct sensitive attributes, and Mean Squared Error (MSE).

$$(MSE) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Mutual Information Reduction: The difference in mutual information between the true and reconstructed values indicates how much information leakage decreases. For efficiency, training time and inference latency are used as metrics to reflect the practical overhead of each defense. The same computational platform will be used for all experiments to ensure consistency. These metrics are summarised in Table 5 for clarity and completeness.

Table 5. Evaluation Metrics.

Category	Metric	Purpose in Evaluation
Classification	Accuracy	Overall correctness
	F1 Score	Balance of precision and recall
	Precision/Recall	False positive/negative rate
Privacy Risk	Inversion Accuracy	Vulnerability to attribute reconstruction
	Mutual Info Reduction	The degree of information leakage mitigated
Regression	MSE	Quality of attribute reconstruction
Efficiency	Training Time	Deployment feasibility
	Inference Latency	Real-time applicability
Regulatory Map	Compliance Alignment	GDPR/HIPAA readiness

Figure 3 represents the architectural layout of the five core defense strategies evaluated in this study. Each subfigure illustrates the key operational flow and intervention point of a specific method. In Figure 3a, Differential Privacy introduces noise during model training to mask sensitive attribute correlations. Figure 3b depicts Federated Learning, where multiple client devices perform local training and only share aggregated gradients with a central server, preserving data locality. Figure 3c shows Hybrid Encryption, which secures data during transmission and storage through ECC-based key exchange and AES encryption. In Figure 3d, Adaptive Noise Injection identifies high-risk features and perturbs them with Gaussian noise before training. Finally, Figure 3e illustrates Ensemble Obfuscation, where predictions from multiple diverse models are combined via a voting mechanism, reducing deterministic leakage and making inversion attacks less effective. Together, these diagrams highlight how each technique targets a specific vulnerability stage in the machine learning pipeline.

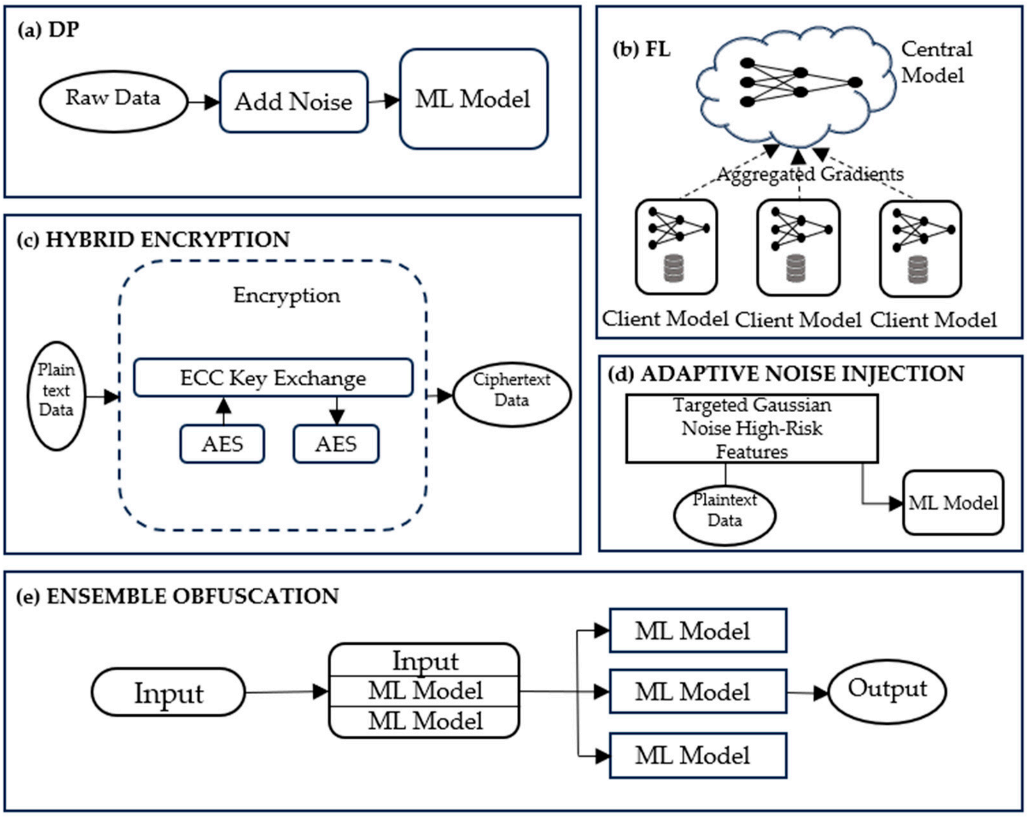

Figure 3. Architecture of five defense strategies. (**a**) Differential Privacy; (**b**) Federated Learning; (**c**) Hybrid Encryption; (**d**) Adaptive Noise Injection: (**e**) Ensemble Obfuscation.

3.5. Experimental Protocol, SOTA Benchmarking, and Complexity Considerations

To ensure rigor, the same random seeds are used in each experiment, and the results are reported in terms of mean and standard deviation. In cases where feasible, the suggested framework is compared against the state-of-the-art (SOTA) techniques that are mentioned in recent works [17,21,42]. Where direct SOTA implementation is not possible (e.g., no code, the dataset is incompatible), the results obtained based on the original publications are provided for comparison. They also include white-box attack experiments in which a partial knowledge of the model parameters is conferred on the adversary to evaluate robustness outside of black-box conditions.

Complexity and Overhead: The time used during training, volume of communication, and latency in differencing are logged for each defense and analyzed. For example, the extra time that federated learning requires and the encryption/decryption steps in hybrid encryption are highlighted.

Reproducibility: A repository containing all code, parameter grid, and configuration files will be publicly released as an additional artifact. This will enable validation of the reported 30–50 percent reductions in inversion accuracy and utility trade-offs.

4. Experiments and Results

4.1. Baseline Vulnerability: Attack Reconstruction Accuracy Without Defense

First, the sensitivity of common machine learning models to model inversion attacks was systematically examined to establish a benchmark. Without employing any protection strategies, Decision Tree, Random Forest, and Multi-Layer Perceptron (MLP) models were trained on three well-known tabular datasets (ADULT, GSS, FTE). These datasets were selected because they are widely used in existing research, making them a relevant standard for comparison with the latest (SOTA) privacy-preserving techniques.

Every experiment was performed using a 70/30 data split, except for the FTE dataset with MLP, which used an 80/20 stratified split to better reflect real-world class distribution. To evaluate stability and address class imbalance, Random Forest was run five times on the ADULT dataset. To manage class imbalance, the SMOTE method was employed in every setup involving imbalanced datasets. These baseline results are shown in Table 6 and Figure 4 and indicate that privacy risks were significant due to the reconstruction of sensitive attributes.

Table 6. Baseline Vulnerability (No Defense Applied).

Model	Dataset	Model Accuracy	Inversion Accuracy	Sensitive Attribute	Preprocessing
Decision Tree	Adult	0.807 ± 0.035	0.76 ± 0.004	Race	Label encoding
Decision Tree	GSS	0.5137 ± 0.0035	0.7157 ± 0.0151	Xmovie	Label encoding
Decision Tree	FTE	0.3505 ± 0.0432	0.7400 ± 0.0389	Infidelity	Label encoding
Random Forest	Adult	0.8279 ± 0.0033	0.8551 ± 0.0113	Race	Label Encoding
Random Forest	FTE	0.7667 ± 0.0551	0.8533 ± 0.0452	Infidelity	Label encoding
Random Forest	GSS	0.6084 ± 0.0032	0.7784 ± 0.0121	Divorce	Label encoding
MLP	GSS	0.6105 ± 0.0227	0.8667 ± 0.03471	Infidelity	One-hot, SMOTE
MLP	FTE	0.7522 ± 0.0747	0.7463 ± 0.1610	Infidelity	One-hot, Label Encoding
MLP	Adult	0.7904 ± 0.0061	0.8722 ± 0.0126	Race	One-hot for target

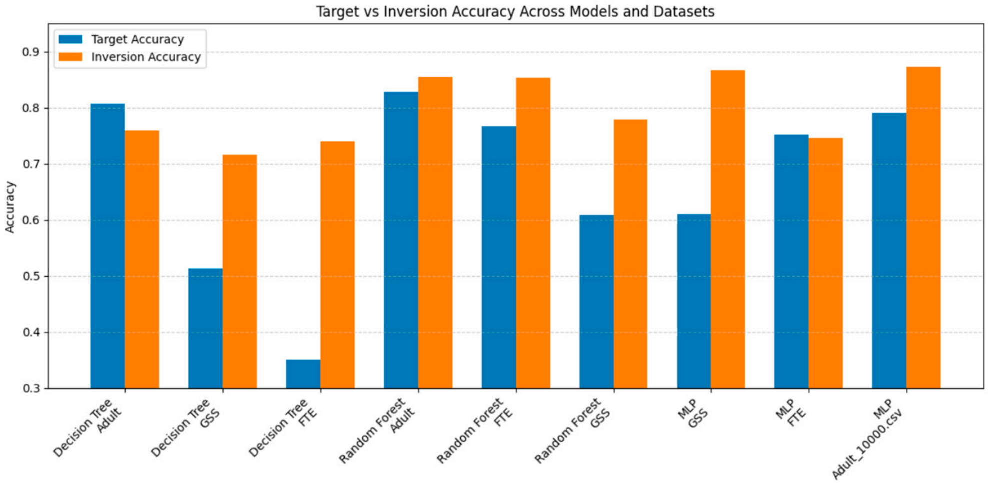

Figure 4. Baseline Vulnerability to Model Inversion Attacks (No Defense).

The outcomes reveal a significant weakness, highlighting the necessity of defensive methods, as shown by the scenario where sensitive features, including race and infidelity, are reconstructed with up to 80 percent accuracy.

4.2. Utility–Privacy Trade-Off: Evaluating the Balance Under Defensive Conditions

This paper examined the concept of privacy utility trade-off, especially in the context of Differential Privacy (DP) via the privacy budget (indicated as 0.96). By raising the magnitude of noise (minor 1/epsilon), the quality of attribute reconstructions fell dramatically, at the expense of functionality as a classifier. Figure 5 illustrates this trade-off by showing the change in both target model accuracy and inversion accuracy across different ε values. Table 7 provides a concrete numerical visualization of the impact of changing the value of (mathematically) to the privacy level as well as utility which shows that it is quite easy to create a setting using the parameters of such policies where several optimal levels of both parameters could be chosen based on the application domain (e.g., in healthcare or general applications):

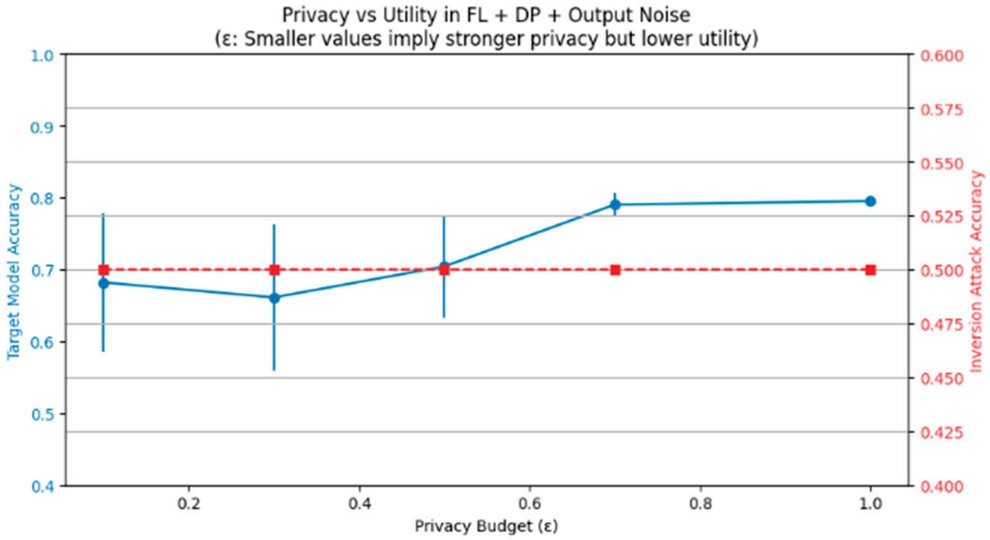

Figure 5. Utility–Privacy Trade-off under a Hybrid Defense using FL and Differential Privacy. The solid blue line (left Y-axis) represents the target model accuracy, while the dashed red line (right Y-axis) indicates the inversion attack accuracy. The privacy budget (ε) controls the trade-off; smaller ε values represent stronger privacy but typically reduce utility.

Table 7. Privacy–Utility Trade-off (Differential Privacy Focus).

ε	Accuracy (Mean ± Std)	F1-Score (Mean ± Std)	Notes on Trade-Off
0.1	59.3% ± 3.6%	62.6% ± 3.3%	Strong privacy, but high utility loss
0.3	71.5% ± 3.5%	72.2% ± 3.1%	Balanced privacy-utility, preferred in sensitive domains
0.5	70.6% ± 4.6%	71.4% ± 4.4%	Balanced but slightly unstable
0.7	78.8% ± 1.6%	77.7% ± 0.7%	Moderate privacy, good utility for general use
1	79.5% ± 0.7%	78.9% ± 1.1%	Low privacy, near-baseline performance

The trade-off clearly indicates diminishing returns in privacy improvement beyond a certain threshold (ε = 0.5).

4.3. Comparative Evaluation Across Defense Mechanisms

A rigorous comparison of five state-of-the-art defense mechanisms was performed: Differential Privacy, Federated Learning, Hybrid Encryption, Adaptive Noise Injection, and Ensemble Obfuscation. The comparison was fair because the same datasets were used, and similar experiments were carried out.

Table 8 presents the results, where Hybrid Encryption offered the most privacy protection, albeit with an increase in computational overhead. Federated Learning [27] represented a good tradeoff between excellent privacy and only a minor tradeoff in accuracy. The

ANI and Ensemble Obfuscation produced good results in performance-driven situations. These trends are also visualised in Figure 6, which compares the target and inversion metrics across all defense strategies.

Table 8. Performance Comparison Across Defense Mechanisms on Adult Dataset.

Defense	Target Accuracy	Target F1	Inversion Accuracy	Inversion F1
Differential Privacy	0.7896 ± 0.0414	0.7849 ± 0.0473	0.4983 ± 0.0176	0.3317 ± 0.0192
Federated Learning + DP	0.7705 ± 0.0312	0.779 ± 0.0122	0.5 ± 0.0112	0.3333 ± 0.0043
Hybrid Encryption (AES + ECC)	0.8505 ± 0.0223	0.8432 ± 0.0218	0.5 ± 0.0415	0.3333 ± 0.0402
Adaptive Noise	0.8490 ± 0.0000	0.8400 ± 0.0000	0.4697 ± 0.0026	0.4165 ± 0.0025
Ensemble Obfuscation	0.8714 ± 0.0000	0.8633 ± 0.0000	0.4589 ± 0.0059	0.4070 ± 0.0053

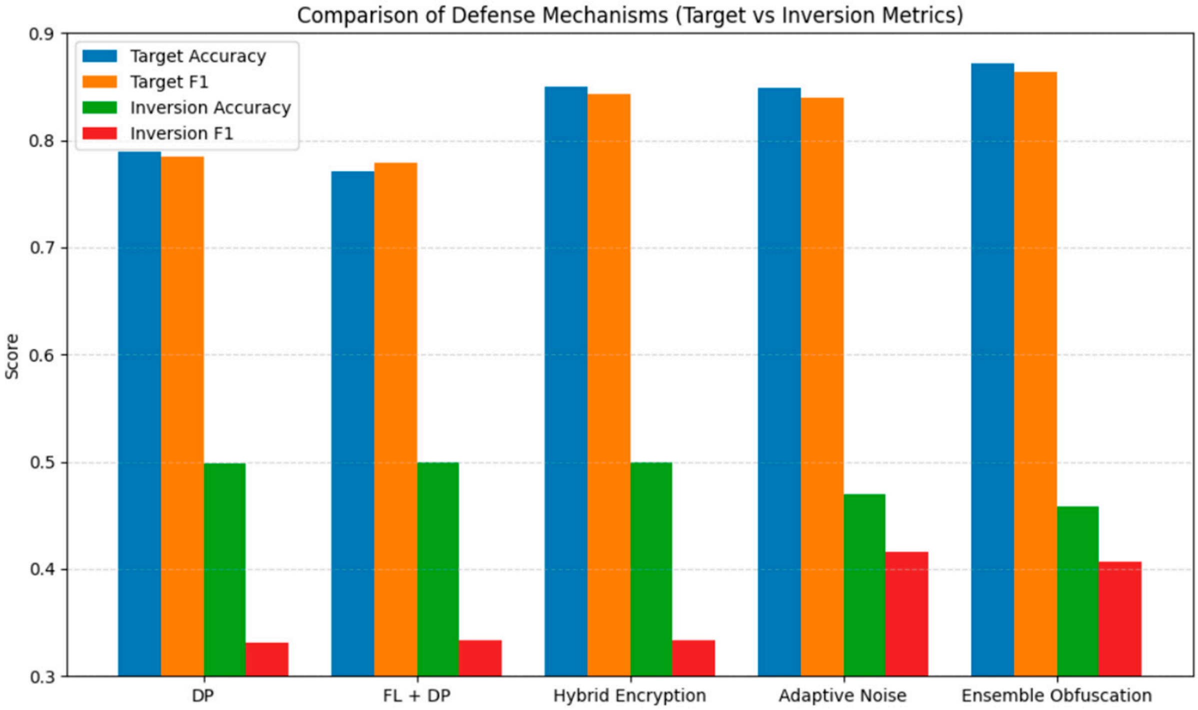

Figure 6. Comparison across Defense Mechanisms.

4.4. Generalization Across Datasets and Domains

The robustness of the results across different domains (Census, Finance, Social Survey) (Table 9) demonstrates their generalizability, as the privacy improvements are moderate to high across all three data sets.

Table 9. Comparison Across Datasets.

Defense	Dataset	Inversion Accuracy (Without Defence)	Inversion Accuracy (With Defence)	Reduction (%)
Differential Privacy	ADULT	87.22%	49.83%	37.39%
	GSS	86.67%	57.36%	29.31%
	FTE	74.63%	58.70%	15.93%

Table 9. *Cont.*

Defense	Dataset	Inversion Accuracy (Without Defence)	Inversion Accuracy (With Defence)	Reduction (%)
Adaptive Noise Injection	ADULT	87.22%	46.97%	40.25%
	GSS	86.67%	65.10%	21.57%
	FTE	74.63%	58.30%	16.33%
Ensemble Obfuscation	ADULT	87.22%	45.89%	41.33%
	GSS	86.67%	55.20%	31.47%
	FTE	74.63%	52.15%	22.48%
Federated Learning + DP	ADULT	87.22%	55.30%	31.92%
	GSS	86.67%	56.50%	30.17%
	FTE	74.63%	60.20%	14.43%
Hybrid Encryption (AES + ECC)	ADULT	87.22%	50.04%	37.18%
	GSS	86.67%	55.50%	31.17%
	FTE	74.63%	53.60%	21.03%

Ensemble Obfuscation produced an average privacy gain that was significantly higher than that of any other method, and Differential Privacy exhibited varying privacy gains across domains, indicating sensitivity to dataset-specific structural correlations.

4.5. Alignment with Data Protection Regulations (GDPR and HIPAA)

Lastly, the compliance of defense strategies with GDPR and HIPAA was clearly analyzed. The Hybrid Encryption demonstrated the greatest regulatory alignment with GDPR Article 32 and the Technical Safeguards as described by HIPAA. The principles of data mineralization in GDPR corresponded with Differential Privacy, and the decentralized handling of data in HIPAA aligned with Federated Learning. Although our technical defenses are aligned with GDPR and HIPAA provisions, this should be considered indicative and does not replace professional legal interpretation. Table 10 provides a structured summary of the mapping between each defense mechanism and relevant GDPR and HIPAA clauses.

Table 10. Defense–Regulation Mapping (GDPR/HIPAA).

Defense Mechanism	Compliance Feature	GDPR Article	HIPAA Clause
Differential Privacy	Data Minimization	Article 5(1)(c)-Principles relating to processing of personal data	§164.502(b)–Minimum Necessary
Federated Learning	Decentralized Data Processing	Article 25(2)-Data protection by design and by default	§164.308(a)(3)–Workforce Security
Hybrid Encryption (ECC + AES)	Secure Data in Transit & Rest	Article 32(1)(a)-Security of processing	§164.312(a)–(e)–Technical Safeguards
Adaptive Noise Injection	Context-Aware Data Perturbation	Recital 78-Appropriate Technical and Organizational Measures	§164.514–De-identification
Ensemble Obfuscation	Risk Reduction via Diversity	Article 25(1)-Data protection by design and by default	§164.306(b)–Security Standards

The compliance of the proposed methods with the regulatory practices also confirms their practical suitability in sensitive areas of application.

5. Limitations and Future Work

While this paper provides a robust and comprehensive analysis of model inversion vulnerabilities and the effectiveness of various defensive strategies, several limitations must be acknowledged to contextualize the findings and inform future research.

5.1. Limitations of the Study

The study focused on well-known tabular models instead of newer examples like gradient-boosted trees, transformers for structured data, or graph networks because these architectures can have different weaknesses and require special protection methods. Another concern relates to the datasets chosen for the study. Although the Adult, GSS, and FTE datasets are different and uniquely organized, they are still carefully crafted datasets with few noisy or disorganized attributes. While the alignment of defenses with GDPR was briefly discussed, the study did not include legal audits, formal risk assessments, or stakeholder impact evaluations.

5.2. Future Scope

Future research should examine a wider variety of machine learning architectures beyond the typical tabular models discussed here. This includes assessing model inversion risks in newer structured-data models such as TabNet and transformers. Using the proposed framework in more fields like social media, e-commerce, and medical records will help confirm its broad applicability. This work does not include a formal analysis of computational complexity. Considering the range of techniques examined, expanding on this aspect could be a meaningful direction for future studies.

Additionally, helping regulators understand and utilize technical barriers to crime should be a top priority. Future efforts should focus on linking legal concepts to specific algorithms. Society benefits when machine learning researchers, lawyers, and officials collaborate to explore techniques for protecting devices and meeting obligations across regions.

6. Conclusions

The study addresses a significant and growing privacy challenge related to machine learning systems involving structured tabular data, specifically the vulnerability to model inversion attacks. Through a systematic analysis and comparison of five prominent defense mechanisms, the study demonstrates that, to a significant extent, privacy risks from model inversion attacks are mitigated without substantially compromising the models' predictive accuracy.

In extensive experiments, defense mechanisms all showed a reduction in success rates of inversion attacks by approximately 30–50 percent, depending on the technique and dataset involved. One of these methods, the Ensemble Obfuscation, was the most successful because it maintained excellent model utility (F1-score of 86%) and significantly lowered inversion accuracy to around 46%. This performance balance suggests that Ensemble Obfuscation is a practical choice in applications where it is crucial to preserve privacy while also maintaining model quality.

Additionally, cross-domain analysis clearly demonstrated the applicability and effectiveness of the proposed defense strategies overall, thereby emphasizing their usability across various real-life situations. Furthermore, strict adherence to leading international data privacy regulations, particularly GDPR and HIPAA, will enhance their functionality and legal viability, especially in sectors like healthcare, finance, and government.

In summary, the work presented in this study establishes a broad, practical, and regulation-compliant method for safeguarding models on sensitive tabular datasets. Going forward, this framework should be expanded to include dynamic IoT systems with real-time threat detection, along with research into adaptive defenses that proactively counter adversarial strategies. As part of our ongoing effort to balance data privacy with technological usefulness, it is essential to maintain a framework that upholds these principles ethically and securely while evolving alongside the rapid advancements in machine learning that impact our most critical fields. To promote reproducibility and encourage further research, all code will be publicly available through a GitHub repository upon acceptance of the paper. The implementation was developed using Python 3.10, with key libraries including PyTorch 2.0, TensorFlow 2.11, Scikit-learn 1.2.2, NumPy 1.24, and Pandas 1.5.3.

Author Contributions: Conceptualization, N.V., R.G., and K.C.C.; methodology, N.V. and K.C.C.; software, N.V.; validation, N.V., R.G., and K.C.C.; formal analysis, N.V. and K.C.C.; investigation, N.V. and K.C.C.; resources, R.G. and K.C.C.; data curation, N.V. and K.C.C.; writing—original draft preparation, N.V.; writing—review and editing, R.G. and K.C.C.; visualization, N.V. and K.C.C.; supervision, R.G. and K.C.C.; project administration, R.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mehnaz, S.; Dibbo, S.V.; Kabir, E.; Li, N.; Bertino, E. Are Your Sensitive Attributes Private? Novel Model Inversion Attribute Inference Attacks on Classification Models. In Proceedings of the 31st USENIX Security Symposium, Boston, MA, USA, 10–12 August 2022; USENIX Association: Boston, MA, USA, 2022.
2. Zhou, S.; Ye, D.; Zhu, T.; Zhou, W. Defending Against Neural Network Model Inversion Attacks via Data Poisoning. *IEEE Trans. Neural Netw. Learn. Syst.* **2025**, *36*, 16324–16338. [CrossRef]
3. Weiss, J.C.; Natarajan, S.; Peissig, P.L.; McCarty, C.A.; Page, D. Machine Learning for Personalized Medicine: Predicting Primary Myocardial Infarction from Electronic Health Records. *AI Mag.* **2012**, *33*, 33–45. [CrossRef]
4. Linden, G.; Smith, B.; York, J. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Comput.* **2003**, *7*, 76–80. [CrossRef]
5. Su, X.; Khoshgoftaar, T.M. A Survey of Collaborative Filtering Techniques. *Adv. Artif. Intell.* **2009**, *2009*, 421425. [CrossRef]
6. Dunis, C.L. *Artificial Intelligence in Financial Markets*; Palgrave Macmillan: New York, NY, USA, 2016.
7. Chan, K.C.; Rabaev, M.; Pratama, H. Generation of synthetic manufacturing datasets for machine learning using discrete-event simulation. *Prod. Manuf. Res.* **2022**, *10*, 337–353. [CrossRef]
8. Rabaev, M.; Pratama, H.; Chan, K.C. Leveraging Synthetic Data and Machine Learning for Shared Facility Scheduling. In Proceedings of the International Conference on Information Technology and Applications, Sydney, Australia, 17–19 October 2024; Springer Nature: Singapore, 2024.
9. Fredrikson, M.; Jha, S.; Ristenpart, T. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, Denver, CO, USA, 12–16 October 2015; Association for Computing Machinery: Denver, CO, USA, 2015; pp. 1322–1333.
10. Yale, A.; Dash, S.; Dutta, R.; Guyon, I.; Pavao, A.; Bennett, K.P. Generation and evaluation of privacy preserving synthetic health data. *Neurocomputing* **2020**, *416*, 244–255. [CrossRef]
11. Zhang, C.; Xie, Y.; Bai, H.; Yu, B.; Li, W.; Gao, Y. A survey on federated learning. *Knowl.-Based Syst.* **2021**, *216*, 106775. [CrossRef]
12. Zhou, Z.; Zhu, J.; Yu, F.; Li, X.; Peng, X.; Liu, T.; Han, B. Model Inversion Attacks: A Survey of Approaches and Countermeasures. *arXiv* **2024**, arXiv:2411.10023. [CrossRef]
13. Gougeh, R.A. How Adversarial attacks affect Deep Neural Networks Detecting COVID-19? *Res. Sq.* 2021, *preprint*. [CrossRef]
14. Kulkarni, Y.; Bhambani, K. Kryptonite: An Adversarial Attack Using Regional Focus. In *Applied Cryptography and Network Security Workshops*; Springer International Publishing: Cham, Switzerland, 2021.
15. Brendel, W.; Rauber, J.; Bethge, M. Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models. *arXiv* **2017**, arXiv:1712.04248.
16. Khamaiseh, S.Y.; Bagagem, D.; Al-Alaj, A.; Mancino, M.; Alomari, H.W. Adversarial Deep Learning: A Survey on Adversarial Attacks and Defense Mechanisms on Image Classification. *IEEE Access* **2022**, *10*, 102266–102291. [CrossRef]

17. Zhang, Z.; Umar, S.; Hammadi, A.Y.A.; Yoon, S.; Damiani, E.; Ardagna, C.A.; Bena, N.; Yeun, C.Y. Explainable Data Poison Attacks on Human Emotion Evaluation Systems Based on EEG Signals. *IEEE Access* **2023**, *11*, 18134–18147. [CrossRef]
18. Zhang, T.; He, Z.; Lee, R.B. Privacy-preserving Machine Learning through Data Obfuscation. *arXiv* **2018**, arXiv:1807.01860. [CrossRef]
19. Li, X.; Qu, Z.; Zhao, S.; Tang, B.; Lu, Z.; Liu, Y. LoMar: A Local Defense Against Poisoning Attack on Federated Learning. *IEEE Trans. Dependable Secur. Comput.* **2023**, *20*, 437–450. [CrossRef]
20. Alzubaidi, L.; Al–Dulaimi, K.; Obeed, H.A.-H.; Saihood, A.; Fadhel, M.A.; Jebur, S.A.; Chen, Y.; Albahri, A.S.; Santamaría, J.; Gupta, A.; et al. MEFF—A model ensemble feature fusion approach for tackling adversarial attacks in medical imaging. *Intell. Syst. Appl.* **2024**, *22*, 200355. [CrossRef]
21. Guo, Y.; Yin, P.; Huang, D. One-Pixel Attack for Continuous-Variable Quantum Key Distribution Systems. *Photonics* **2023**, *10*, 129. [CrossRef]
22. Jingyi, G.; Jianming, W.; Ping, Z. Frequency-guard: Defense against data poisoning attacks to local differential privacy protocols. In Proceedings of the SPIE, San Francisco, CA, USA, 25–31 January 2025.
23. Huang, K.; Ouyang, G.; Ye, Q.; Hu, H.; Zheng, B.; Zhao, X.; Zhang, R.; Zhou, X. LDPGuard: Defenses Against Data Poisoning Attacks to Local Differential Privacy Protocols. *IEEE Trans. Knowl. Data Eng.* **2024**, *36*, 3195–3209. [CrossRef]
24. Hossain, M.T.; Badsha, S.; La, H.; Islam, S.; Khalil, I. Exploiting Gaussian Noise Variance for Dynamic Differential Poisoning in Federated Learning. *IEEE Trans. Artif. Intell.* 2025; 1–17, *early access*. [CrossRef]
25. Rathod, V.; Nabavirazavi, S.; Zad, S.; Iyengar, S.S. Privacy and Security Challenges in Large Language Models. In Proceedings of the 2025 IEEE 15th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 6–8 January 2025.
26. Das, B.C.; Amini, M.H.; Wu, Y. Security and Privacy Challenges of Large Language Models: A Survey. *ACM Comput. Surv.* **2025**, *57*, 152. [CrossRef]
27. Carlini, N.; Tramer, F.; Wallace, E.; Jagielski, M.; Herbert-Voss, A.; Lee, K.; Roberts, A.; Brown, T.; Song, D.; Erlingsson, U. Extracting training data from large language models. In Proceedings of the 30th USENIX Security Symposium (USENIX Security 21), Vancouver, BC, Canada, 11–13 August 2021.
28. Li, H.; Chen, Y.; Luo, J.; Wang, J.; Peng, H.; Kang, Y.; Zhang, X.; Hu, Q.; Chan, C.; Xu, Z. Privacy in large language models: Attacks, defenses and future directions. *arXiv* **2023**, arXiv:2310.10383. [CrossRef]
29. Qu, W.; Zhou, Y.; Wu, Y.; Xiao, T.; Yuan, B.; Li, Y.; Zhang, J. Prompt Inversion Attack Against Collaborative Inference of Large Language Models. In Proceedings of the 2025 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA, 12–15 May 2025.
30. Du, H.; Liu, S.; Zheng, L.; Cao, Y.; Nakamura, A.; Chen, L. Privacy in Fine-Tuning Large Language Models: Attacks, Defenses, and Future Directions. In *Advances in Knowledge Discovery and Data Mining*; Springer Nature: Singapore, 2025.
31. Petrov, I.; Dimitrov, D.I.; Baader, M.; Müller, M.N.; Vechev, M. Dager: Exact gradient inversion for large language models. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 87801–87830.
32. Lukas, N.; Salem, A.; Sim, R.; Tople, S.; Wutschitz, L.; Zanella-Béguelin, S. Analyzing Leakage of Personally Identifiable Information in Language Models. In Proceedings of the 2023 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA, 22–25 May 2023.
33. Zheng, H.; Hu, H.; Han, Z. Preserving User Privacy for Machine Learning: Local Differential Privacy or Federated Machine Learning? *IEEE Intell. Syst.* **2020**, *35*, 5–14. [CrossRef]
34. Bos, J.W.; Halderman, J.A.; Heninger, N.; Moore, J.; Naehrig, M.; Wustrow, E. Elliptic Curve Cryptography in Practice. In *Financial Cryptography and Data Security*; Springer: Berlin/Heidelberg, Germany, 2014.
35. Alves, T.; Das, R.; Morris, T. Embedding Encryption and Machine Learning Intrusion Prevention Systems on Programmable Logic Controllers. *IEEE Embed. Syst. Lett.* **2018**, *10*, 99–102. [CrossRef]
36. Panzade, P.; Takabi, D.; Cai, Z. Privacy-Preserving Machine Learning Using Functional Encryption: Opportunities and Challenges. *IEEE Internet Things J.* **2024**, *11*, 7436–7446. [CrossRef]
37. Bokhari, M.U.; Shallal, Q.M. A review on symmetric key encryption techniques in cryptography. *Int. J. Comput. Appl.* **2016**, *147*. [CrossRef]
38. Li, Y.; Liu, F. Adaptive Gaussian Noise Injection Regularization for Neural Networks. In *Advances in Neural Networks—ISNN 2020*; Springer International Publishing: Cham, Switzerland, 2020.
39. Phan, N.; Wu, X.; Hu, H.; Dou, D. Adaptive Laplace Mechanism: Differential Privacy Preservation in Deep Learning. In Proceedings of the 2017 IEEE International Conference on Data Mining (ICDM), New Orleans, LA, USA, 18–21 November 2017.
40. Tan, Y.X.M.; Elovici, Y.; Binder, A. Adaptive Noise Injection for Training Stochastic Student Networks from Deterministic Teachers. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Virtual, 10–15 January 2021.

41. Conti, M.; Vinod, P.; Vitella, A. Obfuscation detection in Android applications using deep learning. *J. Inf. Secur. Appl.* **2022**, *70*, 103311. [CrossRef]
42. Zhao, J.; Chen, Y.; Zhang, W. Differential Privacy Preservation in Deep Learning: Challenges, Opportunities and Solutions. *IEEE Access* **2019**, *7*, 48901–48911. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.