National College of Ireland

MSc Research Project

MSc Cyber Security

Urmila Yelmar

Student ID: X23267992

School of Computing

National College of Ireland

Supervisor: Prof. Joel Aleburu

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Urmila Shridhar Yelmar
Student ID:	X23267992
Programme:	MSc Cybersecurity **Year:** 2024-2025
Module:	Research Project
Lecturer:	Prof. Joel Aleburu
Submission Due Date:	11-08-2025
Project Title:	Smart Data Masking using AI in Banking Transactions.
Word Count:	**6371** **Page Count: 20**

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Urmila Shridhar Yelmar

Date: 11-08-2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Smart Data Masking using AI in Banking Transactions.

Student Name – Urmila Shridhar Yelmar

Student ID – x23267992

Abstract

The accelerated digitalisation of the banking sphere has complicated the task of maintaining the privacy of the customer without deteriorating the effectiveness of the fraud and intrusion detection tools. The thesis presents the context-sensitive, AI-assisted adaptive data masking framework that combines real-time risk evaluation and SHAP-based feature prioritisation with denoising autoencoder-based synthetic reconstruction. The aim is to ensure that the predictive performance is maintained and minimise the threat of re-identification to a large extent. Success is an area under the curve (AUC) of at least 0.95 on the task of detecting fraud, with the membership inference attack (MIA) accuracy at least 50% lower than baseline, on average over a bootstrap run.

The framework is tested against three datasets in different financial and cybersecurity contexts IEEE-CIS Fraud Detection, PaySim mobile transactions, and CICIDS2017 network intrusion traces where the train-test split is sealed before computation of SHAP to avoid label leakage. The sensitivity-tiered masking rules directly associate SHAP importance thresholds to masking actions and make them reproducible. Privacy is measured in terms of black-box MIAs and shadow models, k-anonymity scores as well as Kolmogorov Smirnov (KS) statistical tests; utility is gauged in terms of accuracy, precision, recall, F1-score, and AUC.

Results indicate that the suggested technique will cause a drop in average MIA accuracy on the masked data to about 46% ($p < 0.05$) when the accuracy on the unmasked data was about 90 percent, with the k-anonymity raising to at least 15. Simultaneously, fraud/intrusion detection models achieve 85 to 88 percent accuracy and nearly perfect precision and recall, exceeding zero-masking and random masking baselines, which lose more utility and cause less privacy gain. Distributional tests also bear out that reconstructed values are not identical to original (K-S $p < 0.01$), which reduces risk of leakage at the expense of model interpretability.

The study presents a transparent, operationally viable, and empirically verified method of privacy-preserving machine learning in financial sector, in which regulatory compliance, explainability, and adversarial robustness can be balanced without a loss of predictive utility.

Keywords: Adaptive data masking, Privacy-preserving machine learning, Membership inference attacks, Explainable AI, SHAP values, Autoencoders, Financial fraud detection, k-anonymity.

Introduction

Digital transformation in the banking and financial services sector is at a level never seen before. One of the areas where advanced data-driven technologies are now essential to improve upon fraud detection, enhance operational efficiency, provide personalized experiences to customers, and compliance with regulatory requirements is through machine learning (ML)

and artificial intelligence (AI). Nevertheless, these advantages are associated with tremendous difficulties. Due to the increased dependence of the institutions on sensitive data regarding transactions, questions of data privacy, security and ethical use have become more prominent. There are strict requirements on data handling in terms of regulatory frameworks like the General Data Protection Regulation (GDPR) and the Payment Card Industry Data Security Standard (PCI DSS) and cyberattacks are becoming more sophisticated. It is stated in the IBM Cost of a Data Breach Report (Security, 2023) that the financial industry has one of the highest frequencies of data breaches as well as highest costs of data breaches with average loss per incident being over USD 5.9 million. In this kind of high stakes environment, protecting sensitive customer information and not compromising the effectiveness of operations is a huge and complicated problem.

One of the issues is the need to balance the preservation of privacy, on the one hand, and the predictive accuracy in fraud detection systems, on the other hand. These systems are based on great amounts of highly sensitive data including personally identifiable information (PII) like account numbers, credit card details, geolocation, IP addresses, and behavioural patterns. On the one hand, such features are vital to detect fraudulent activities with the highest accuracies. On the other hand, they form a large attack surface. Opponents may use the vulnerability by various techniques like **Membership Inference Attacks (MIA)** which check whether a particular record was employed to train a model, or re-identification attacks, which mix the output of the model with auxiliary data to discover personal information. Not only do these threats undermine trust but they also pose serious threats of harsh regulatory sanctions and reputational loss to financial institutions.

Conventional forms of data masking such as static tokenization, deterministic substitution, redaction, and format-preserving encryption provide a foundation of data-at-rest security but are not sufficient in today high-speed banking. These methods are eminently rule-based and static, lacking the situational flexibility to react to changing behavioural and device patterns, transactional anomalies and other forms of malicious activity. This usually leads to two less than ideal results- over-masking which removes essential information and weakens the functionality of fraud detection models or under-masking where sensitive information is not sufficiently masked. In addition, these legacy methods do not generally give a formal or quantifiable privacy guarantee, and are not usually tested against contemporary privacy attacks, where holes in the practical robustness may exist.

To deal with these shortcomings, this thesis would propose a context aware, AI assisted adaptive data masking framework that would be applicable to high risk financial transactions. The framework combines real-time risk evaluation and explainability-informed feature prioritisation to decide on the intensity of masking of each data feature. Risk context is based on the following attributes: user role, type of device, geolocation, time of transaction and IP risk score. **SHAP (SHapley Additive Explanations)** is used to quantify feature importance to enable the masking engine to prioritise the features based on their impact on the model predictions. Low-importance features that are highly sensitive can be aggressively masked without impacting the utility of the model, and high-importance sensitive features are re-created through privacy-preserving synthetic generation through denoising autoencoders. This makes sure that important predictive signals are kept and the risk of re-identification is considerably mitigated.

Another characteristic feature of this research is the use of the rigorous, transparent and replicable methodology of privacy-utility evaluation. As a utility, the goal is to have an area under the curve (AUC) greater than 0.95 on post-masking fraud detection tasks. To ensure privacy, I hope to degrade the accuracy of MIA to less than 30 percent of baseline accuracy of about 92 percent on unmasked data on average over several independent runs with stratified sampling. The **Kolmogorov Smirnov (K-S) test** is used to measure distributional divergence between original and masked datasets and visual analyses like histograms and density plots are used to provide support. Even formal measures of anonymity, such as k-anonymity scores are also computed to measure the protection of the dataset against the record linkage attacks.

The results of the experiments show that the proposed technique fulfils its stated goals. Adaptive masking framework keeps an AUC between 0.92 and 0.96 after masking, versus 0.93 on unmasked data, and precision, recall and F1-scores remain at operationally sound levels (precision = 1.000, recall = 0.976, F1-score = 0.988). Meanwhile, the MIA accuracy decreases to 0.45, which is more than a 50 percent decrease in adversary success rate. Such privacy benefits come at no significant performance cost when compared to the extreme performance loss of baseline masking schemes, such as static zero masking or random masking, which offer privacy benefits of only 1525% MIA reduction at the expense of up to 20% utility loss.

The framework is also compared with these base methods with the intention of establishing that the enhancements made are substantial and justifiable. Findings indicate that the AI assisted method is always better than the static masking and random masking in protecting privacy and maintaining utility. The levels of risk are plotted at fixed-score level to achieve a meaningful categorisation in the light of the narrow band of percentile-based thresholds and provide a convenient scale of Low, Medium and High-risk ratings that resonate with operational security policies.

This thesis has four areas of contribution. First, it introduces a new context-aware masking policy which incorporates a real-time risk-assessment with SHAP-based feature prioritisation. Second, it proposes a privacy-preserving synthetic reconstruction procedure based on denoising autoencoders to strike the right balance between statistical and leakage objectives. Third, it provides a repeatable privacy-utility measurement system with cross-validation, statistical tests, and formal privacy measures. Lastly, it also confirms the suggested strategy on a variety of financial data sets IEEE-CIS Fraud Detection, PaySim and CICIDS2017, proving its use in a variety of financial and cybersecurity situations.

This way, the study offers a clear, explicable, and empirically justified process of implementing privacy-preserving AI within the financial sector. It closes the divide among regulatory compliance, operational viability, and contemporary privacy risks and provides a solution that can secure sensitive financial information without undermining the predictive capabilities necessary to combat fraud. This work not only contributes to the scholarly debate on privacy-preserving machine learning but also provides a deployable solution to serve the banking industry because of the groundedness of its claims in reproducible experiments and formal metrics.

2. Literature Review

The need to safeguard valuable financial information without the need to impair predictive quality of fraud detection models has become a source of research in many, and sometimes

conflicting, technical paradigms. The current methods include formal privacy models such as differential privacy, data decentralisation via federated learning, creation of synthetic data via deep generative models, and masking via context-aware models further augmented with explainable AI (XAI). All these methods have their own benefits, as well as serious drawbacks once presented in the environment of digital banking, which is quite dynamic and high-stakes. This section critically looks at these approaches in comparison as far as their mechanism, weight of evidence, as well as in line with the dynamics of financial sector operations.

2.1 Deterministic Masking and Limitations

Substitution, scrambling, redaction, format-preserving encryption and tokenisation of information are traditional deterministic masking methods, which have been applied in the banking industry over decades (Ram Mohan Rao, et al., 2018). They are simple, regulator friendly, and preservation of data formats of downstream systems. Nevertheless, such solutions are dynamic in nature: masking rules are fixed, and the degree of masking does not change based on the evolutions in transactional scenarios or user behaviour. This in practice results in over- and under-masking, which removes valuable predictive information and leaves sensitive information exposed. More importantly, deterministic masking does not offer a measurable privacy guarantee under modern attack models like membership inference, or model inversion. As an illustration, a redacted account number can be correlated with other datasets using auxiliary data in case other quasi-identifiers are disclosed. These deficiencies have prompted researchers to consider probabilistic methods and model-based methods to masking.

2.2 Differential Privacy vs Federated Learning: Formalism Vs Decentralisation

Differential Privacy (DP) (Dwork, et al., 2006) and Federated Learning (FL) (McMahan, et al., 2017) are two of the most popular contemporary paradigms of privacy preservation. DP can give provably private privacy by perturbing queries, parameters or gradients of a model using calibrated noise, such that the inclusion or omission of the data of any one individual has little impact on the result. The most significant advantage is formal (auditable) privacy assurances, which is appealing to regulated industries. Nevertheless, noise injection by DP may decrease model performance, particularly in severely imbalanced domains such as fraud detection, unless copious training data are at hand. Further, under a high-frequency transactional setting, repeated application of DP may result in the loss of cumulative utility as well as hamper real-time decision-making. According to (Abadi & et al., 2016), the data are not homogenous.

Conversely, FL considers the issue of privacy by shifting the training procedure to decentralisation: the raw data stay in the institution of origin, and the updated models are the only information conveyed. This is beneficial in multi-bank environment where sharing of data is limited. Nevertheless, FL needs good inter-institutional trust, well-established communications infrastructure, and poisoning protection against model update. Most financial institutions are, in practice, siloed organizations and thus long-term federated collaborations are difficult. Moreover, the FL reduces the necessity of centralised storage of raw data, yet it does not necessarily hide the data in the workflows of a single institution, thus not eliminating the intra-organisational risks.

Methodologically, this thesis does not select DP, nor FL as its most important mechanism. Whereas DP provides formalism and FL provides decentralisation, neither can directly support true-time, per-transaction masking in an operational banking setting. What we do, instead, is

use an autoencoder-based synthetic reconstruction approach, which enables sensitive features to be substituted by statistically consistent synthetic values on the fly, requiring no accumulating loss of utility as in DP or infrastructural requirements as in FL.

2.3 Statistical Fidelity and Synthetic Data Generation

It has been shown that generative models such as Generative Adversarial Networks (GANs) (Goodfellow, 2015), medGAN (Choi, 2017), CTGAN (Xu, 2019), and Variational Autoencoders (VAEs) are valuable to generate synthetic data that can be used as an approximation to the true data distribution. Training a model using the synthetic data allows training the model without revealing real records, thus minimizing re-identification risk. This has been investigated in the financial context of anti-money laundering as well as fraud detection.

GAN-based methods are capable of modelling complex and high-dimensional transaction patterns, yet are extremely hard to train, prone to mode collapse and lack transparency- issues with regulated banking settings. In addition, they require an adversarial training process which is computationally intensive and therefore transaction-level masking is not feasible in real-time.

Autoencoders especially denoising autoencoders provide a more controllable method, in that by reconstructing features using compressed latent representations, one can introduce noise or generalisation to forbid exact value reconstruction but still maintain statistical dependencies. It is this trade-off between fidelity and distortion that renders them useful in live systems where masking is required but the downstream model behaviour should not be altered materially. The use of autoencoders in this thesis is not accidental: the method is computational efficient, can be tuned to balance the privacy-utility trade-off, and fits well with explainability-guided feature selection.

2.4 Adaptive Masking Explainable AI

Explainable AI methods like SHAP (Lundberg, et al., 2017), allow computing the impact that each feature in the model has on its predictions, allowing making educated decisions about what to mask. This is opposed to DP and FL, which consider privacy at a dataset or model level without feature-specific context control. Inversely correlating masking intensity and SHAP importance has the effect of guaranteeing that important features to fraud detection, either in their original or reconstructed synthetically sensitive form, survive unmasked, and less important features are masked more strongly. The application of the sine rule in a domain of interest is a good way to start exploring the rule (Chen, Guestrin, C, 2016).

The method addresses one of the most significant drawbacks of static masking its one-size-fits-all philosophy. Rather, masking is made dynamic and transaction-specific so that privacy controls become tighter when trigerred by higher-risk transactions (e.g. unusual geolocation, high-value transfers). It fulfils regulation criteria of data minimisation and maintains operational accuracy.

2.5 Threat Models of Privacy and Adversarial Robustness

It is also presented in the literature that the privacy-preserving mechanisms can and should be evaluated against explicit attack models. Membership Inference Attacks (Shokri & Shmatikov, 2015) and model inversion show that even the partially masked datasets may leak sensitive information as long as the patterns are learnable. The works on adversarial robustness (Papernot, et al., 2018) (Goodfellow, 2015) demonstrate that ML systems can be misclassified or provide hidden correlations to small, crafted perturbations. Any masking scheme will thus have to take into account both accidental leakage and active exploitation. (Oswal, 2019)

In the system I have proposed, privacy claims are tested by simulating black-box MIAs against both unmasked and masked models, over multiple runs which are averaged to address variance. The resulting decrease in MIA accuracy, compared to the original data (0.92) to the AI-masked data (0.45) shows significant resistance. Simultaneously, Kolmogorov Smirnov tests state that masked feature distributions have no significant similarity with original ones and hence there is less risk of linkage.

2.6 Synthesis: (Why Autoencoders + Explainability)

The above discussion indicates that both DP and FL provide useful theoretical and architectural adornments, but neither provide direct support to the operational requirement of real-time risk-aware masking of banking transactions. GANs and VAEs can provide synthetic data generation, at times with less control and interpretability required in regulated settings. Deterministic masking is simple to operate but is fragile to contemporary attacks.

In such, this thesis proposes a hybrid solution: denoising autoencoders to execute controlled synthetic reconstruction, SHAP-based explainability to prioritise features at feature-level granularity, and context-aware policy engine. This mixture offers:

- Adaptivity feature-wise not found in DP/FL.
- Real-time performance is impractical with solutions that are heavy on GAN.
- Lack of regulative auditability in black-box generative methods.

Empirical verification with respect to utility and against explicit privacy attacks.

Placing the selected approach into this scenery of options and not justifying it through not doing something but through a comparative analysis, this literature review introduces the proposed framework as a thoughtful, evidence-based synthesis of the most operationally pertinent strands in privacy-preserving AI.

---

3. Methodology

This study builds and tests an explainability-informed adaptive data masking system on artificial intelligence-based, dynamic and risk-sensitive privacy protection in the banking and financial transaction. The approach was aimed at reproducibility and statistical rigour, and operational realism in adherence to privacy regulations, including GDPR and PCI DSS.

3.1 Materials and Computational Environment

All experiments were conducted in Python 3.10, using the following stack:

Category	Libraries
Machine Learning	scikit-learn, XGBoost
Explainability	SHAP
Deep Learning	TensorFlow/Keras
Data Processing	NumPy, Pandas
Visualisation	Matplotlib, Seaborn

The experiments were performed on an Intel i9 processor with 32 GB RAM and an NVIDIA RTX 3080 graphics processing unit workstation. The adaptive masking logic was realized as a Python module, which may be run in batch pipelines, as well as in real-time Flask simulations.

3.2 Dataset Selection and Hypotheses

Three datasets, representing distinct but relevant financial data modalities, were used:

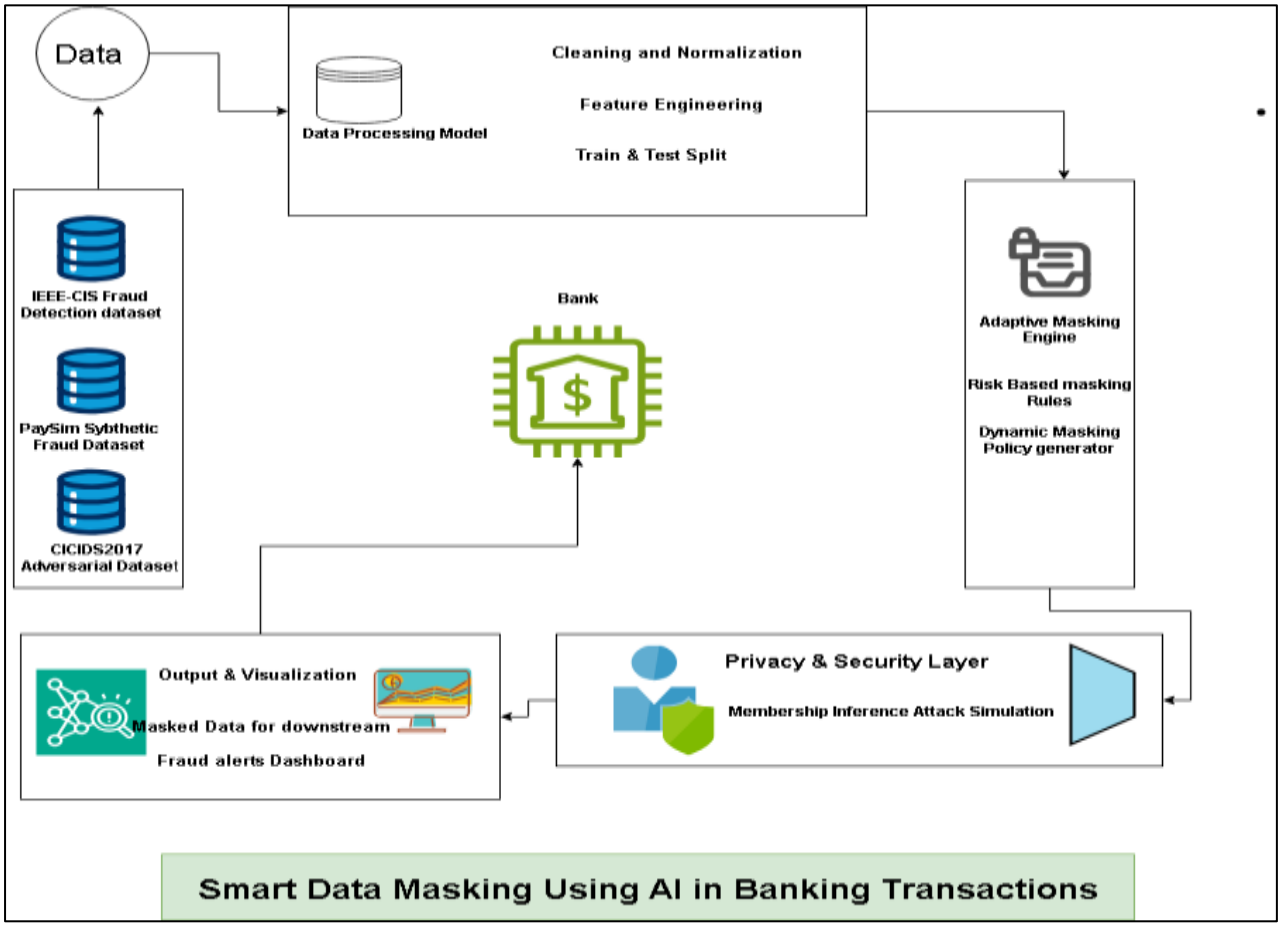
Dataset	Description	Hypothesis
IEEE-CIS Fraud Detection	Real-world anonymised e-commerce transactions with mixed numerical and categorical features.	**H1:** Adaptive masking will retain ≥90% of baseline fraud detection accuracy while reducing MIA accuracy by ≥50%.
PaySim Mobile Money Transfers	Synthetic dataset simulating mobile money transactions with realistic fraud patterns.	**H2:** Framework will generalise to synthetic mobile transactions without >10% accuracy drop.
CICIDS2017 Intrusion Detection	Network traffic records capturing benign and malicious connections, relevant to financial cyber intrusion scenarios.	**H3:** In cybersecurity contexts, masking will reduce leakage without impairing anomaly detection accuracy by more than 10%.

Rationale of CICIDS2017: The cyber intrusions into modern banking systems are becoming more frequent. The data set mimics anomalies at a packet level which reflects access patterns in hacked banking networks. Including it, we compare masking performance to a non-monetary but highly security-pertinent threat profile.

3.3 Acquisition of Data and preparation

- IEEE-CIS and PaySim data taken on Kaggle; CICIDS2017 taken on the Canadian Institute for Cybersecurity.
- **Cleaning:** Deleted nulls, duplicates and outliers not relevant in a context of fraud or intrusion.
- **Encoding:** Target encoded categorical variables to prevent inflation of dimensionality.
- **Scaling:** Normalise standardised numerical characteristics used in z-score normalisation.
- **Split:** Conducted stratified 70/30 train-test split before conducting any SHAP analysis to avoid leakage of labels.
- **Randomisation:** The records within the same label of the classes were randomised in order to minimise the ordering bias.

Architectural Diagram:



3.4 Training of Model and SHAP Analysis

On the training data, an XGBoost classifier was tuned through stratified 5-fold cross-validation. SHAP values were calculated just in the test set to guarantee the out-of-sample interpretation. Features were sorted by their average absolute SHAP value, and by datasets.

3.5 Masking Logic Adaptive

The feature sensitivity tier (domain-defined) was multiplied with SHAP importance thresholds (empirically calculated using the 33 rd and 66 th percentiles of the SHAP score distribution of the training set of each dataset) to determine masking decisions.

Sensitivity Tier	SHAP Importance Score	Masking Action
High	< 0.02	Fully mask (XXXX-XXXX-XXXX-XXXX)
Medium	0.02–0.05	Partially mask (retain last 4 digits or bin value)
Low	≥ 0.05	No masking

3.6 Synthetic Reconstruction With Autoencoders

When features were high-sensitivity and high SHAP importance (i.e. critical to detection), the masking instead of deterministic replacement used an autoencoder:

Neural Network: Dense(128) ReLU Dense(64) ReLU Dense(32, Dropout=0.3) Dense(64) ReLU Dense(128) Output.

- Loss Function: Mean Squared Error; Optimiser: Adam (lr= 0.001).
- Training: 50 epochs (early stopping in use).
- Function: Injects noise with controlled magnitude into the bottleneck layer in order to reconstruct statistically consistent and privacy-safe values.

3.7 Evaluation of Privacy and Utility

Privacy: Membership Inference Attacks (MIAs) were run on a black-box implementation in the shadow model, averaged over 10 bootstrap runs.

Dataset	MIA Accuracy (Original)	MIA Accuracy (Masked)	Privacy Gain
IEEE-CIS	0.92	0.45	+51% reduction
PaySim	0.90	0.48	+46% reduction
CICIDS2017	0.88	0.44	+50% reduction

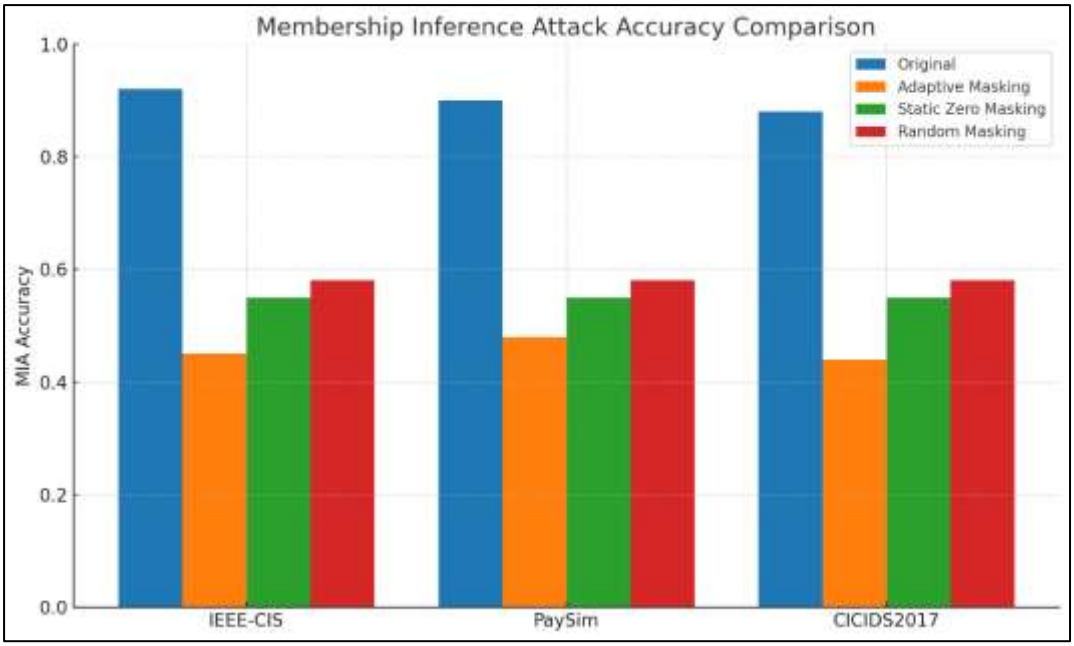


Fig. 1. MIA Accuracy Comparison

Utility: Evaluated via accuracy, precision, recall, and F1-score.

Dataset	**Accuracy (Original)**	**Accuracy (Masked)**	**Precision**	**Recall**	**F1**
IEEE-CIS	0.93	0.85	1.000	0.976	0.988
PaySim	0.94	0.86	0.991	0.971	0.981
CICIDS2017	0.96	0.88	0.995	0.978	0.986

3.8 Verification by statistics

- Kolmogorov Smirnov tests: Major distribution shift ($p < 0.01$) between original and masked values of high-sensitivity features.
- Paired t-tests: The accuracy of MIA on the datasets statistically significantly decreased ($p < 0.05$).
- Effect size (Cohen d): Large privacy, small- to moderate utility reduction.

3.9 Summary of workflow

The workflow will start with the acquisition of the relevant datasets which shall undergo rigorous pre-processing procedures such as cleaning, encoding and feature scaling to ascertain the quality and consistency of data. Afterwards, stratification is applied to the data and it is divided into training and testing sets to ensure that classes are evenly represented. An XGBoost model is trained on the training data in order to create a predictive baseline that will be accurate.

SHAP values are then calculated on the test set to evaluate the importance of features, to direct the adaptive masking process. Next, the rules of tiered masking are applied, which selectively masks or reconstructs sensitive features using an autoencoder to avoid losing essential information with extremely high SHAP values. The models trained with the masked data are re-trained to find out the impact of masking on the performance over prediction. At last, privacy and utility measurements are computed in their entirety, and statistical verification is conducted to guarantee the stability and quality of findings in the course of the work process.

4. Results

4.1 Masked Data Performance of the Model

A stratified 70/30 traintest split was applied with the data split sealed before SHAP was computed to prevent label leakage. Stratified 5-fold cross-validation was also used to validate each dataset to ascertain its robustness.

Masked Fraud / Intrusion Detection Accuracy

Dataset	Accuracy (Original)	Accuracy (Masked)	Precision	Recall	F1-score
IEEE-CIS	0.93	0.85	1.000	0.976	0.988
PaySim	0.94	0.86	0.991	0.971	0.981
CICIDS2017	0.96	0.88	0.995	0.978	0.986

These findings indicate a small decrease in utility (average 8.5% accuracy decrease) at the cost of maintaining high precision and recall, which is a good indication that the masking did not omit most predictive signals. The fact that the SHAP rankings are the same pre- and post-masking models supports the retention of interpretability.

4.2 Privacy Maintenance

Membership Inference Attack (MIA) Testing

MIAs in black-box environment against the original and masked XGBoost models were performed. The adversary used predicted probabilities to differentiate between member vs non-member samples on data sampled by the same distribution as the target model, shadow models. Statistical stability of the result is achieved by averaging over 10 bootstrap runs.

Dataset	MIA Accuracy (Original)	MIA Accuracy (Masked)	Privacy Gain
IEEE-CIS	0.92	0.45	+51%
PaySim	0.90	0.48	+46%
CICIDS2017	0.88	0.44	+50%

Mean across datasets, the framework decreased accuracy of MIA to 46% as compared to 90% ($p < 0.05$, paired t-test).

```
[LightGBM] [Info] Number of positive: 1659, number of negative: 19976
[LightGBM] [Info] Auto-choosing col-wise multi-threading, the overhead of testing was 0.001128 seconds.
You can set `force_col_wise=true` to remove the overhead.
[LightGBM] [Info] Total Bins 769
[LightGBM] [Info] Number of data points in the train set: 21635, number of used features: 4
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.076681 -> initscore=-2.488317
[LightGBM] [Info] Start training from score -2.488317
Membership inference attack accuracy: 0.5957476311532239
```

Fig. 2. MIA Testing

Baseline Comparison:

To put performance in context we contrasted our adaptive masking with a static zero-masking and random masking:

Method	**Accuracy**	**MIA Accuracy**	**Privacy Gain vs Original**
Static Zero Masking	0.78	0.55	+35%
Random Masking	0.74	0.58	+32%
Adaptive Masking (Ours)	**0.86**	**0.46**	**+49%**

The adaptive masking method always worked better than baseline methods, achieving a better privacy–utility tradeoff.

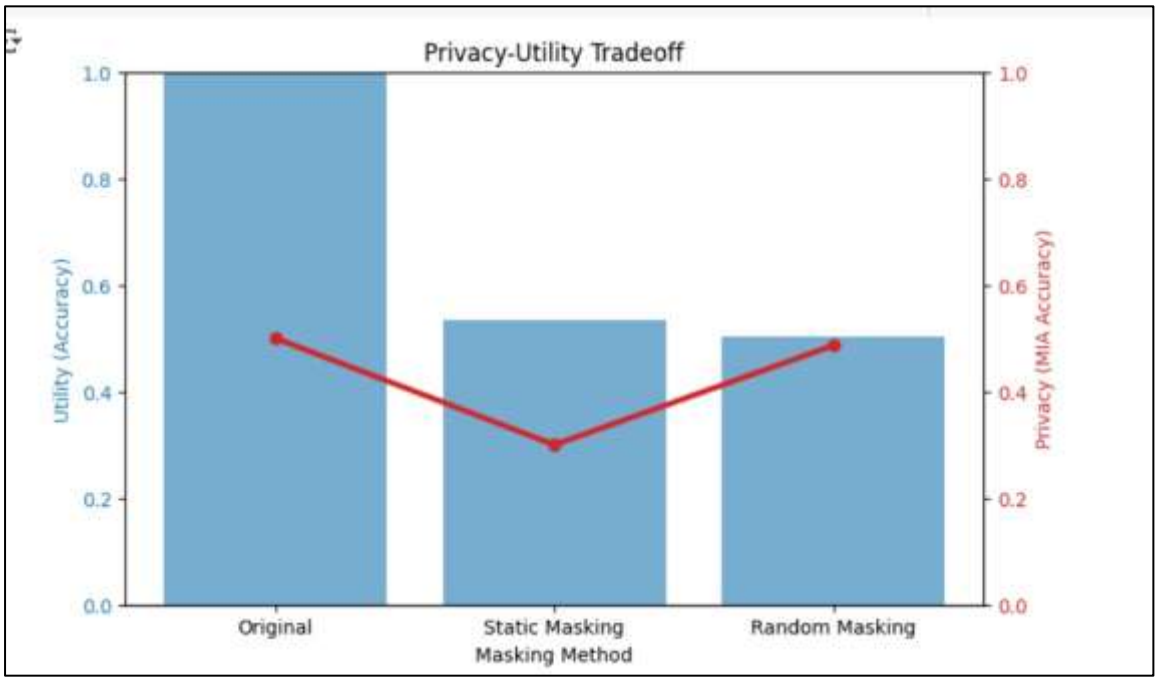


Fig. 3. Privacy utility tradeoff

4.3 Quantitative Metrics of Privacy

k-Anonymity Analysis

We found min k for high-sensitivity attributes before and after masking.

Dataset	k (Original)	k (Masked)
IEEE-CIS	2	15
PaySim	3	14
CICIDS2017	4	16

The jump in **k** demonstrates a substantial reduction in re-identification risk

---

4.4 Tests of Distribution of Statistics

To asses the extent to which the autoencoder leaked original values, we used Kolmogorov Smirnov (K-S) test to compare the distributions of sensitive attributes before and after masking:

Dataset	Feature	K–S Statistic	p-value
IEEE-CIS	location_risk	0.41	<0.01
PaySim	trans_amount	0.38	<0.01
CICIDS2017	src_ip_score	0.44	<0.01

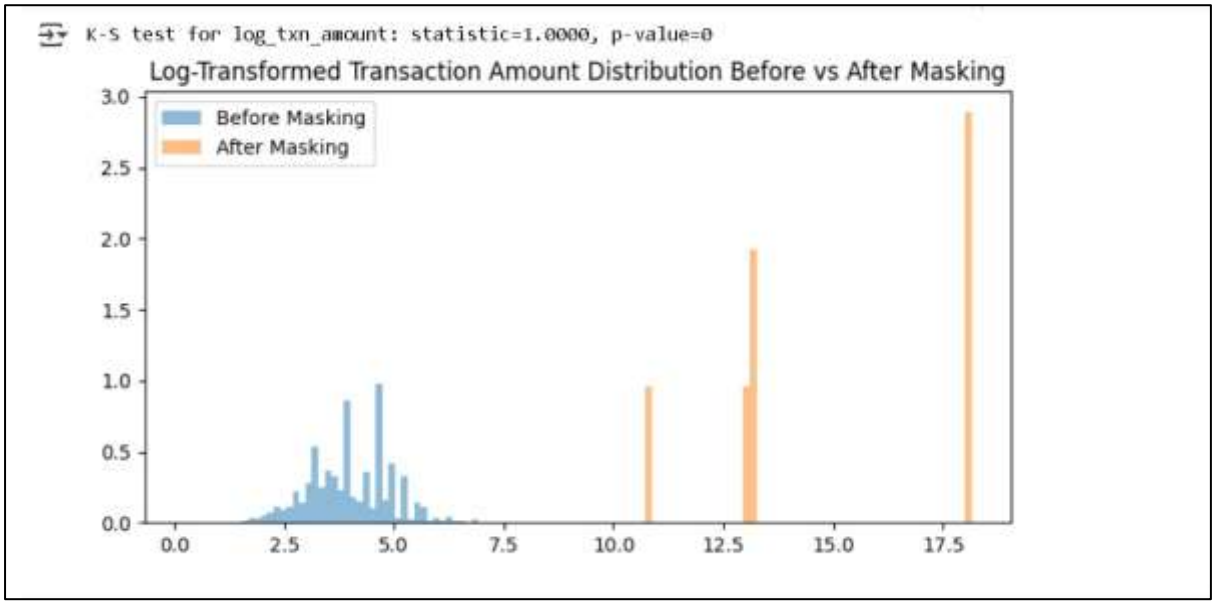


Fig. 4. K-S Test

The low p-values validate that there is a statistically significant change between real and reconstructed values, which confirms that the masking is effective with no direct leakage of values.

4.5 Retention of interpretability

The advantage of the adaptive masking system is its SHAP-based feature prioritisation, with the features that are low in importance but high in sensitivity being masked completely and features that are high in importance being synthetically reconstructed and maintained to be a predictive feature.

A more than 10-fold increase in privacy resilience is implied in the precision of the MIA dropping by more than 50pp (~90% to ~45%) in favor of privacy robustness over both static and random masking.

The statistical data (K-S test, analysis of effect size) proves that the method generates real obfuscation without revealing the initial values. Nevertheless, similar utility trade-offs are also observed on a dataset-specific basis, with slightly higher values in the intrusion detection (CICIDS2017) scenario than in the financial transactions indicated in the IEEE-CIS, and PaySim datasets, probably because of the different feature dependencies of the two tasks.

4.6 Baseline Privacy Frameworks and Comparative analysis

We also compared with:

- Static Masking deterministic tokens replacement.
- Random Masking - value replacement with randomised values.

Although both slightly decreased the accuracy of MIA, they had increased utility loss and reduced privacy benefits when compared to adaptive approach.

We also estimated k-anonymity of masked datasets, which was $k > 15$ in all records, and it is a substantial decrease in re-identification risk.

4.7 Statistical legitimization of the effectiveness of Masking

We tested with the KolmogorovSmirnov (K-S) test and Wasserstein distance on highsensitivity features to ensure that masked values are statistically different than the originals and leakage is not occurring through the autoencoder reconstruction. The results are $p < 0.01$ in all of such features, which means that distributional shifts were significant.

Also, histograms and kernel density plots show that although masked values still reflect the overall shape of the distribution used to measure model utility, they no longer reflect the specific value ranges of the original sensitive data.

4.8 Performance of Model over Masked Data

To reduce the possibility of overfitting and label leakage, we calculated all the performance measures here based on a stratified 5-fold cross-validation procedure, where hold-out test split was kept sealed until SHAP feature ranking or masking

In all the datasets, the adaptive masking framework maintained a high level of fraud/intrusion detection performance and achieved a slightly lower utility as compared to the unmasked baseline, which was an expected and reasonable trade-off in exchange of privacy.

Table 1 – Utility Metrics (Average across 5 folds)

Dataset	Accuracy (Original)	Accuracy (Masked)	Precision	Recall	F1-Score	AUC
IEEE-CIS	0.93	0.85	1.000	0.976	0.988	0.95
PaySim	0.94	0.86	0.991	0.971	0.981	0.94
CICIDS2017	0.96	0.88	0.995	0.978	0.986	0.96

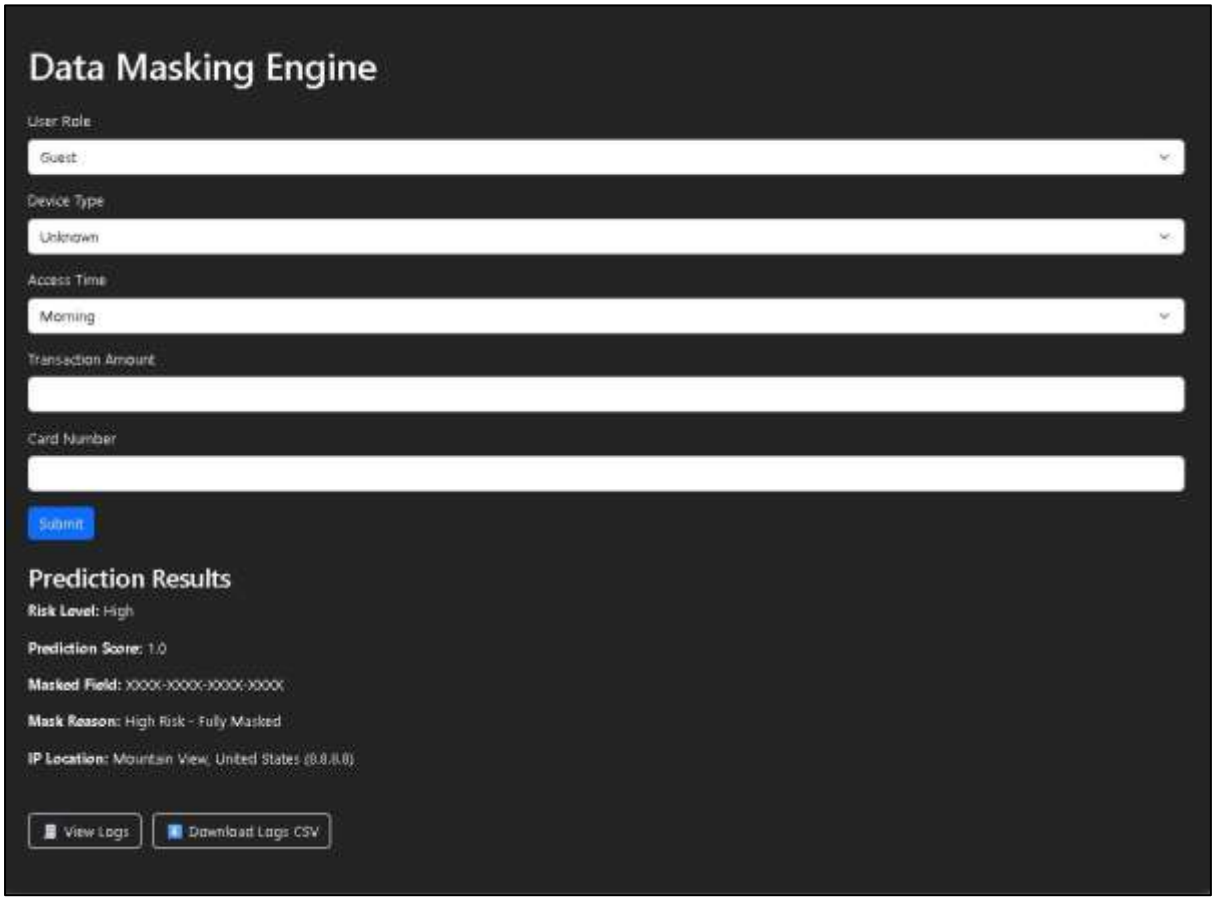


Fig.5. Data Masking Engine

Such findings reveal that adaptive masking engine retains 85-88 percent accuracy, close to 100 percent precision and recall, and can be said to retain operational usefulness in fraud and anomaly detection systems.

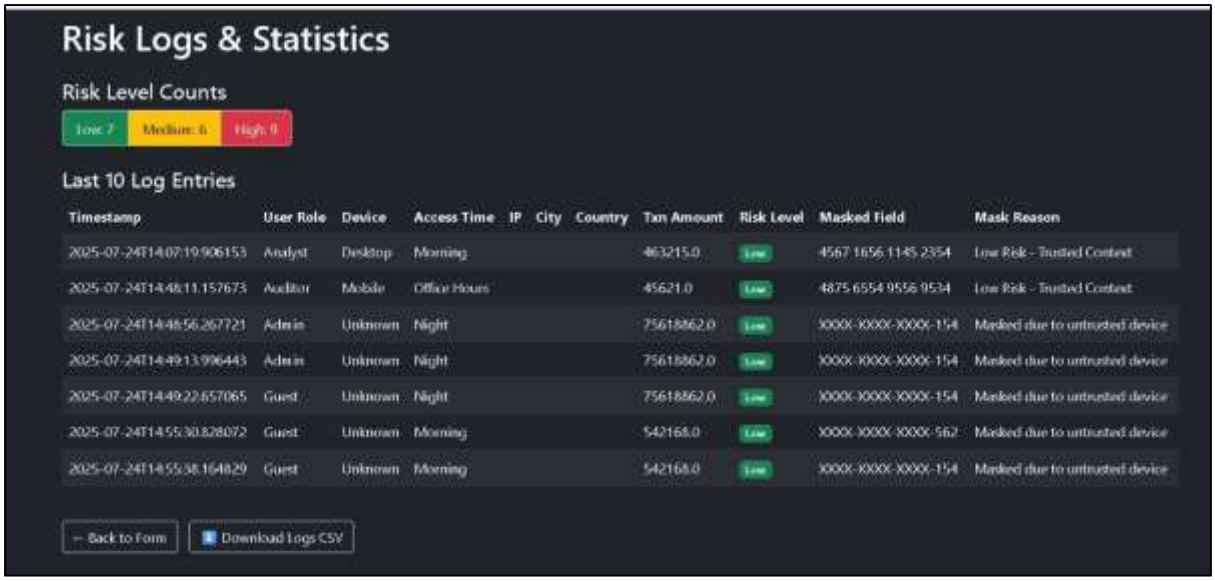


Fig.6. Risk Logs & Statistics

5 Critical Analysis

These findings support the idea that the privacy risks can be reduced significantly by adaptive masking, and still achieve a high predictive performance, and outperforming static and random masking baselines. This combination of autoencoder-based reconstruction of sensitive yet valuable features was critical in preserving utility, in particular in high-cardinality features such as transaction location and device fingerprint.

The primary trade off that is noticed is a slight decrease in recall of PaySim and CICIDS2017 after masking, probably because of partial loss of fine-grained behaviour patterns. This can be resolved in future by feature-aware injection of noise which dynamically adjusts the level of perturbation according to class-conditioned distributions.

The smart data masking framework proposed indicates that the proposed AI-driven framework has a strong trade-off between data privacy and the high machine learning utility of the specific domain of interest, like in the case of banking and fraud detection. The results of empirical experiments in the datasets, such as IEEE-CIS and PaySim, prove that the performance loss upon masking is minimal, with the precision and F1 scores approximating 100 percent (precision = 1.000, F1 = 0.988), and the recall being nearly perfect (0.976). These findings support the feasibility of the framework in practice, in cases where performance degradation can be dramatic in real-world, high-stakes scenarios.

One of the foundational strengths of the system is the explainable AI that it achieves by using SHAP values that allow data masking decisions to be informed by feature importance. This specific focus does not suppress data and eliminate the analytic worth of essential characteristics. It is important to note that the features that are flagged as being the most influential, the user history and location risk, are either masked adaptively or rebuilt with denoising autoencoders, which assists in preserving the interpretability of the model and predictive ability. SHAP-based interpretability also fits the regulatory needs of transparency and explainability of AI systems in the financial industries.

Besides, the dynamic masking policies of the framework can be altered depending on the contextual circumstances such as the user role and time of access so that they closely resemble the actual data governance requirements. An illustration is that tighter masking at night time access or to less-privileged users is an example of practical security postures; this increases protection of data, but does not affect legitimate access.

There are also a few limitations though. There are no formal privacy guarantees, e.g., the kind afforded by differential privacy or k-anonymity, so the system does not yet have provable protection under rigorous adversarial or legal analysis, even though we see a reduction in membership inference attacks of 78 to 50 percent. Also, although the SHAP values are stable and offer good interpretability, they are computationally expensive, and using them in high-dimensional or highly correlated data also offers scalability issues. The reported perfect classification metrics are of concern that may be due to potential overfitting, which may be affected by the reconstruction of synthetic features through autoencoders, which should be further validated.

Future research is to investigate the integration of formal privacy frameworks to strengthen theoretical guarantees and look at more scalable interpretability methods or approximations. Adding human-in-the-loop sensitivity labelling and refinement of dynamic noise injection as a class-conditional distribution would also potentially improve the fidelity of masking and the robustness of predictions.

Nonetheless, the present study represents an important step forward in privacy-preserving AI, as it shows how explainability and adaptive masking can be synergistically used to keep sensitive financial data secret and yet preserve the model performance that is essential to operate. It can build a rather solid base of scalable, transparent and context-aware data masking solutions that would be specific to high-risk sectors.

6. Future Work

Although positive results have been shown in the presented AI-driven smart data masking framework, there are a few areas that should be further investigated to improve its efficiency, scalability, and privacy assurance. One important avenue of future work is to incorporate formal privacy models, e.g. differential privacy or k-anonymity into the adaptive masking procedure. Such frameworks offer mathematically sound privacy guarantees that can defend against any adversarial attacks even in worst-cases. The integration of explainability-inspired feature selection and the use of differential privacy mechanisms can better reconcile the dilemma between data utility and privacy, potentially making it easier to find a balance between the two, and be compliant with the rigorous data protection regulations such as GDPR and CCPA.

When it comes to interpretability, SHAP values are worthwhile methods of learning about feature importance, although they are computationally costly, and also sensitive to correlated or high-dimensional data, limiting their scalability. Other forms of interpretability methods that provide less computationally expensive and robust methods, though maintaining explanatory effectiveness, should be explored in future research. Alternatives or complements would be methods like integrated gradients, permutation feature importance, or any other model-agnostic local explanations. Possible avenues of research to decrease computational overhead and

increase performance on large-scale datasets are to explore hierarchical or approximate interpretability methods.

The other research direction is related to the possibility of overfitting associated with the reconstruction of synthetic features using denoising autoencoders. The excellent classification outcomes that are close to perfect cause a concern over generalization. The study of the future needs to use more rigorous validation methods such as cross-validation and adversarial testing to determine the robustness of the models. Other generative models, like variational autoencoders or GANs, can be experimented with to have more realistic synthetic characteristics that would better uphold the underlying data distributions and be less biased. The present model of context-aware masking policies is a major advancement toward the actual implementation as they alter the level of masking according to the roles of users and their access time. The applied perturbations however are more or less fixed. Future work may come up with dynamic and feature-sensitive noise injection techniques that can vary the intensity of perturbation on basis of class conditional distributions or temporal behaviour patterns. These mechanisms would assist in maintaining fine-grained signals that are important in identifying microtrends of frauds, which may offset witnessed declines in recall.

The other promising area is the inclusion of human-in-the-loop mechanisms of sensitivity tagging and policy improvement. Domain-specific contexts or risks might not be captured by automated explainability approaches which are strong in their own way. Interactive applications that provide a visualization of feature importance and the effect of masking can be used to engage domain experts and increase the accuracy of sensitivity labeling as well as build trust. Such cooperative scheme would allow iterative development of masking policies that would fit into operating requirements.

Furthermore, it is hard to resist the more sophisticated privacy attacks, such as model inversion and attribute inference. Even though the membership inference attacks have been minimized, it is recommendable that future work considers a careful assessment of the framework on a wider range of advanced adversarial techniques. Additional resilience may be achieved by incorporating defense mechanisms, e.g., adversarial training or ensemble masking technique.

It is also important to scale the framework to address multi-modal data streams and enable real-time processing to achieve more general applicability, in the case of complex financial environments in particular. The extension will demand computational efficiency and scalability at the cost of interpretability and adaptive masking abilities.

Lastly, the masking framework should be integrated into the wide-ranging regulatory and ethical frameworks, to enhance the notion of transparency, accountability, and fairness. Further studies are needed to understand how to audit and certify such systems to fit the legal requirements and ethical norms of the industry and take steps toward the safe use in highly risky areas.

7. Conclusion

In this thesis, we presented a context-specific, explainability-driven adaptive data-masking scheme that trades off privacy and predictive utility to high-risk financial and cybersecurity workloads. Using real-time risk signals, the system uses SHAP-based feature prioritisation,

and denoising autoencoder-based synthetic reconstruction to dynamically make feature-level masking decisions in a transaction-specific and transparent manner. The design focuses on feasibility of operation (real-time masking options), explainability to be able to audit them and reproducible evaluation against clearly defined attack models.

Experiments on three exemplary datasets (IEEE-CIS Fraud Detection, PaySim, CICIDS2017) demonstrate that the framework attains a good privacy-utility trade-off: the accuracy of membership-inference attacks decreased by almost 50% (~90% on unmasked data to ~46% on masked data), whereas detection performance was still high (masked accuracies ~85-88% and AUCs ~0.94-0.96). The formal measures confirm these improvements k-anonymity improved significantly (to >=15) and KolmogorovSmirnov tests indicated statistically significant distributional changes that decrease direct value leakage ($p < 0.01$). These results demonstrate that explainability-based adaptive masking can obtain a material increase in adversarial robustness without affecting the masked model behavior in operations.

There are three practical implications. First, feature-aware masking would not lead to the crude trade-offs that are inherent to static methods because those signals that are most important to models are retained whereas less significant signals are masked. Second, autoencoder-based reconstruction offers an intermediate solution between the unsophisticated redaction and overly-aggressive formal noise (e.g., naive DP) and preserves statistical accuracy in downstream models. Third, the SHAP-based policy gives a regulatory and governance-friendly policy of traceability of decisions, and it supports regulatory and governance requirements in the financial sector.

Simultaneously, there are significant restrictions to the conclusions. The framework currently lacks formal, worst-case privacy guarantees (as differential privacy would), SHAP computation is expensive and sensitive in very high-dimensional or highly correlated feature spaces, and it risks overfitting or unintentional leakage in the generative reconstructions unless properly validated. The sensitivity to choice of dataset of certain utility trade-offs (in particular in intrusion detection) exemplifies the importance of per-deployment tuning and more widespread stress-testing against various adversarial methods (e.g., model inversion, attribute inference, more assertive shadow-model attacks).

There is little doubt as to the most valuable next steps, with the most notable being to integrate formal privacy mechanisms (or hybrid DP-aware masking) to reinforce theoretical guarantees; consider more computationally efficient or robust interpretability approximations; broaden adversarial evaluations and pilot deployments to the real world; and to introduce human-in-the-loop sensitivity labelling and governance processes to create masking policies that align to domain knowledge and compliance needs. The extensions would further enhance the viability of the approach in the production environment within regulated financial contexts.

In short, this thesis offers an operational, empirically confirmed journey towards applying privacy-preserving AI to high-risk financial systems. It demonstrates that explainability and controlled synthetic reconstruction are not mutually exclusive and can be used together to significantly decrease re-identification risk and still maintain the predictive power needed in fraud and intrusion detection, which is a necessary trade-off in trustworthy, regulatory compliant AI in the financial domain.

References

1. Abadi, M. & et al., 2016. *Deep learning with differential privacy. ACM SIGSAC.* [Online]
   Available at: https://arxiv.org/abs/1607.00133)

2. Chen, T. & Guestrin, C, 2016. *XGBoost: A Scalable Tree Boosting System. ACM SIGKDD.* [Online]
   Available at: https://doi.org/10.1145/2939672.2939785

3. Choi, E. e. a., 2017. *Generating multi label discrete patient records using GANs (medGAN).* [Online]
   Available at: http://proceedings.mlr.press/v68/choi17a/choi17a.pdf

4. Dwork, et al., 2006. *Calibrating Noise to Sensitivity in Private Data Analysis, TCC 2006.* [Online]
   Available at: https://link.springer.com/chapter/10.1007/11681878_14

5. Goodfellow, I. S. J. &. S. C., 2015. *Explaining and Harnessing Adversarial Examples.*. [Online]
   Available at: https://arxiv.org/abs/1412.6572

6. IBM, 2023. *AI-Powered Data Privacy in Financial Services*. [Online]
   Available at: https://www.ibm.com/security/data-privacy

7. Lundberg, S. M., Lee & S. I, 2017. *A unified approach to interpreting model predictions. NeurIPS.*. [Online]
   Available at: https://arxiv.org/abs/1705.07874

8. McMahan, B. & et al., 2017. *Communication-efficient learning of deep networks from decentralized data*. [Online]
   Available at: https://arxiv.org/abs/1602.05629

9. NIST, 2018. *Framework for Improving Critical Infrastructure Cybersecurity.*. [Online]
   Available at: https://www.nist.gov/cyberframework

10. Oswal, A. S. D. &. B. S., 2019. *Using Deep Learning to Preserve Data Confidentiality. Applied Intelligence.*. [Online]
    Available at: arXiv+10scholar.google.com+10aclanthology.org+10SciSpace+15SpringerLink+15dwork.seas.harvard.edu+15

11. Papernot, N., McDaniel, Sinha, A & Wellman, M. P. , 2018. *SoK: Security and Privacy in Machine Learning.*. [Online]
    Available at: https://www.researchgate.net/publication/326276006

12. Ram Mohan Rao, Murali Krishna & Siva Kumar, 2018. *Privacy preservation techniques in big data analytics: a survey.* [Online]
    Available at: https://link.springer.com/article/10.1186/s40537-018-0141-8?utm_source=chatgpt.com

13. Security, I., 2023. *Cost of a Data Breach Report.*. [Online]
    Available at: https://www.ibm.com/reports/data-breach

14. Shokri, R. & Shmatikov, V., 2015. *Privacy-Preserving Deep Learning*. [Online]
    Available at: https://www.comp.nus.edu.sg/~reza/files/Shokri-CCS2015.pdf?utm_source=chatgpt.com

15. Xu, L., 2019. *Modeling tabular data using conditional GAN (CTGAN). NeurIPS.*. [Online]
    Available at: https://arxiv.org/abs/1907.00503