Privacy Sensitivity Analysis in (ϵ, δ)-Differentially Private Deep Learning for Fraud Detection

by

Albert Sallés Torruella

This thesis has been submitted in partial fulfillment for the degree of Master of Science in Artificial Intelligence

in the
Faculty of Engineering and Science
Department of Computer Science

March 2025

Declaration of Authorship

This report, Privacy Sensitivity Analysis in (ϵ, δ)-Differentially Private Deep Learning for Fraud Detection, is submitted in partial fulfillment of the requirements of Master of Science in Artificial Intelligence at Munster Technological University Cork. I, Albert Sallés Torruella, declare that this thesis titled, Privacy Sensitivity Analysis in (ϵ, δ)-Differentially Private Deep Learning for Fraud Detectionand the work represents substantially the result of my own work except where explicitly indicated in the text. This report may be freely copied and distributed provided the source is explicitly acknowledged. I confirm that:

- This work was done wholly or mainly while in candidature Master of Science in Artificial Intelligence at Munster Technological University Cork.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at Munster Technological University Cork or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this project report is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

MUNSTER TECHNOLOGICAL UNIVERSITY CORK

Abstract

Faculty of Engineering and Science
Department of Computer Science

Master of Science

by Albert Sallés Torruella

Fraud detection is crucial for the financial sector, protecting against illicit activities that can lead to significant financial losses. Despite the advancements in anomaly detection, the application of Artificial Intelligence models requires access to sensitive financial data, raising concerns about privacy and data protection. Differential Privacy (DP) provides a promising solution by introducing controlled noise to protect data confidentiality. However, the impact of DP parameters, particularly dataset size and the privacy loss parameter δ, on model privacy and utility remains an open research question. This study investigates the sensitivity of dataset size and δ in (ϵ, δ)-Differential Privacy within deep learning-based fraud detection. A Long Short-Term Memory (LSTM) model will be trained on differentially private datasets, and privacy will be evaluated using Membership Inference Attacks to assess data leakage risks. Additionally, the study will compare the impact of DP on deep learning models versus traditional machine learning models to identify potential differences in privacy-utility trade-offs. By systematically analysing the effects of dataset size and δ on privacy, this research aims to provide valuable insights into optimising differentially private fraud detection models for real-world applications.

Contents

Abbreviations

DP	**D**ifferential **P**rivacy
ML	**M**achine **L**earning
DL	**D**eep **L**earning
LSTM	**L**ong **S**hort-**T**erm **M**emory
SVM	**S**upport **V**ector **M**achine

Chapter 1

Research Proposal

1.1 Research Context

The research context is explained in three different parts. The motivation highlights the importance of the problem faced in the research. Then, the background section describes some important concepts to understand the proposal. And finally the state-of-the-art solutions are explained in the related work section.

1.1.1 Motivation

Artificial Intelligence has significantly impacted the domain of anomaly detection, caused by the high amount of data available and the increasing complexity of detecting irregular patterns on different fields. In the financial sector, fraudulent transactions detection is a well-known problem, deeply explored through the application of Machine Learning (ML) and Deep Learning (DL) algorithms. However, this approach requires authorisation to handle confidential financial data. This presents a challenge in maintaining data privacy, which certain regulations, such as EU *General Data Protection Regulation*, aim to address by setting strict standards. Consequently, anonymisation-based strategies have gained importance to develop privacy-preserving and compliant processes.

1.1.2 Background

Differential Privacy

Differential Privacy (DP) is a data privacy-preserving technique for protecting individual privacy from sensitive datasets. Its goal is that the outcome of any analysis about a

dataset must be the same regardless of whether any single individual's data point is included. This ensures a high level of privacy to any individual's information [1].

Formal Definition: A function F gives (ϵ, δ)-differential privacy if for all neighbouring datasets D_1 and D_2, which differ by only one element, and all $S \subseteq Range(F)$,

$$P[F(D_1) \in S] \leq e^{\epsilon} \cdot P[F(D_2 \in S] + \delta, \quad \epsilon, \delta > 0$$

In this definition, there are two privacy parameters:

- ϵ denotes the privacy loss parameter. Smaller values of ϵ correspond to stronger privacy guarantees.
- δ is a small parameter which allows for a slight probability of the guarantee being exceeded. This may provide better utility for certain scenarios [2].

Membership Inference Attack

In a black box setting, a Membership Inference Attack exploits the observation that machine learning models often behave differently on the data that they were trained on versus the data that they are presented for the first time. Thus, the attacker's objective is to determine whether a record was part of the target model's training dataset.

Shokri et al. (2017) [3] proved this attack effective by creating shadow models to behave the same way as the target model. To mitigate this attack, they mention that if the training is differentially private, the probability of producing a given model from a training dataset that includes a particular record is similar to the probability of producing such model without that record, thus making differential privacy ideal against these attacks.

Assessing the privacy through training shadow models is quite inefficient, so Salem et al. (2018) [4] made the observation that using the original model's predictions on the target points is sufficient to deduce their membership. Hence, no training of any shadow model that approximates the original model's behavior is required which makes the attack more efficient. In this setting, the original model under attack kind of serves itself as a "shadow model" that approximates its own behavior perfectly. This is how TensorFlow Privacy's Membership Inference Attack has been implemented [5], which yields a score of the confidence of this attacker that a particular sample is a member of the training set. However, it is mentioned that this is not necessarily a probability.

1.1.3 Related Work

With the advance in credit card technology, more features have been developed over the years, making the payment processes more accessible and comfortable for the user. Maes et al. (2002) [6] outlined the problems of credit fraud and the features that a detection system should have. In this case, a Bayesian Network was proposed over an Aritificial Neural Network because of its performance. Over the years, however, more Machine Learning models have emerged which performed considerably better than simple shallow ANN. For example, Randhawa et al. (2018) [7] provided a novel approach in detecting fraud cases in credit cards by combining several tree-based models into a majority voting system. The choice of using tree-based models became really important because of the intrinsic interpretability they offer. Because of the lack of interpretability, DL models were left behind in real-world scenarios even though their performance may be on par with ML models, if not better. Raghavan et al. (2019) [8] conducted a comparison in which a CNN model outperformed other ML approaches, concluding that with the rise of available data and computation power, DL models became more successful. Since then, other DL approaches have achieved higher scores, even combining different DL models like LSTM by Mienye et al. (2023) [9] with the help of good data preprocessing techniques. Btoush et al. (2023) [10] published a systematic review on over 180 fraud detection research articles showcasing the recent popularity of DL models and the need for further research on CNN and LSTM models.

Machine Learning models have become more accurate at detecting fraud and anomalies on imbalanced datasets. Nevertheless, credit card data may expose users' sensitive information and Luo et al. (2023) [11] exposed the research gap between financial technology and the privacy protection industry, highlighting the potential of Differential Privacy as a solution to preserve data utility while protecting private information on credit card data. Ruan et al. (2019) [12] was one of the first to apply DP on a custom fraud detection system to assess the privacy concern. Then, Cai et al. (2020) [13] applied DP to a SVM model to predict individual credit card score and Basu et al. (2021) [14] used DP on a financial text classification task with a DL model. More recently, Perez et al. (2023) [15] designed a collaboratively framework for fraud detection which leveraged differential privacy to avoid sharing sensitive data, successfully comparing the trade-off between privacy and utility after applying different values of ϵ in DP.

It is not clear how to choose a good value for ϵ. In the literature, algorithms have been evaluated with ϵ ranging from from 0.01 to even 7 with little to no justification. Hsu et al. (2014) [16] calculated a range of acceptable values for ϵ and δ in (ϵ, δ)-differential privacy. Using N as the data size and X as the record space, which refers to the total

number of possible unique records that can exist in the database, the lower and upper bounds for ϵ are calculated as follows:

$$\frac{1}{N} \leq \epsilon \leq \max\left(\ln(0.1 \cdot |X|), \ln\left(\frac{|X|-1}{|X|(1-0.1)}\right)\right)$$

In the case of δ, however, it is only mentioned that non-private mechanism are the same as $(0, \frac{1}{N})$-differential private. Therefore, for a more reasonable guarantee, we require

$$0 < \delta \ll \frac{1}{N}$$

In the differential privacy literature, the privacy parameter δ has not been deeply studied. As we have seen, dataset size plays a crucial role in (ϵ, δ)-differential privacy, yet most studies focus on balancing privacy and utility within a fixed dataset. This overlooks the fact that real-world applications, such as fraud detection, handle datasets that constantly change in size. As a result, the impact of dynamic dataset variations on privacy and utility remains an open research question. Addressing this gap requires adaptive DP mechanisms that can adjust to evolving data distributions while ensuring effective fraud detection and strong privacy guarantees.

Mienye et al. (2024) [17], after showing the robustness of DL models in fraud detection systems, described the challenges when developing deep learning-based credit card fraud detection systems. In their work, two main concerns were discussed: privacy and interpretability. Due to their complexity, DL models are often considered black-box systems, making it difficult to understand their decisions. However, interpretability is crucial for ensuring reliability, compliance, and user trust in fraud detection systems.

This lack of interpretability also raises the question of whether DP affects DL and traditional ML models differently. Since DP introduces noise to protect privacy, its impact on model performance may vary depending on the model's complexity. DL models, with their multi-layered architectures, may react differently to DP noise compared to ML models, potentially leading to greater performance degradation. Investigating these differences is essential for designing effective privacy-preserving fraud detection models that maintain both accuracy and transparency.

Based on this lack of research on the sensitivity of δ and the dataset size in the scenario of $(\epsilon, \delta)-$differential privacy on a deep learning-based fraud detection systems, this project proposes applying (ϵ, δ)-Differential Privacy to a public fraud detection dataset [18] and analyse the impact in the privacy using different values of δ and differences in the data size by performing a membership inference attack on the DL model. Subsequently, the

same impact will be assessed in a traditional ML scenario to see whether there is a difference in using a deep neural network.

1.2 Research Aim

The aim of this project is to analyse the sensitivity of the dataset size and δ on the privacy of a trained model on (ϵ, δ)-differential privacy dataset, using both a traditional ML model and a deep neural network.

1.3 Research Objectives

1. To investigate how different dataset sizes affect the privacy of the model with a fixed ϵ value.
2. To study the sensitivity of δ on the privacy of the model using membership inference attack.
3. To assess whether differentially private datasets affects differently traditional ML and DL models.

1.4 Research Methodology

In this research, one dataset will be used: European Credit Card Fraud Detection [18]. The Credit Card Dataset contains transactions made by credit cards in September 2013 by European cardholders, which is a highly imbalanced dataset with only 0.172% fraudulent transactions. This dataset is used in many research papers [17] [8] [9] [11].

To address the data imbalance issue, the dataset will be processed according to Alamri and Ykhlef's survey [19], where they employed different sampling techniques in credit card fraud detection tasks and the main technique explored was Synthetic Minority Oversampling Technique (SMOTE). In the study, in fact, they identified hybrid sampling methods as more efficient in handling the imbalance issues, while noticing that oversampling techniques can lead to overfitting and undersampling can discard essential samples.

The first step will be to train a LSTM neural network model for the fraud detection task. The choice of this model is based on benchmarks and surveys provided by recent papers [17] [15] [10]. These papers offer a comparison between different neural network

architectures in this paticular task and LSTM has thrived because of the history knowledge that the model is capable to retain and the link that exists between prediction outputs and historical input. The LSTM architecture is designed to to learn through long-term relations, which has shown great performance in this task. The actual deep neural network will consist of two LSTM layers connected to a final dense layer before the output [17]. Additionally, dropout will be used to avoid overfitting of the minority task. The architecture can be visualised in the following diagram and table.

FIGURE 1.1: LSTM Architecture Diagram: The data flows from the input layer through two LSTM layers with dropout layers in between, followed by a dense layer and finally the output layer.

Layer (type)	**Output Shape**	**Param #**
lstm_1 (LSTM)	(None, 10, 64)	24,064
dropout_1 (Dropout)	(None, 10, 64)	0
lstm_2 (LSTM)	(None, 32)	12,416
dropout_2 (Dropout)	(None, 32)	0
dense_1 (Dense)	(None, 1)	33

TABLE 1.1: Summary of the LSTM model layers showing the output shape and the number of parameters for each layer.

During the training phase, hyperparameter tuning will be performed to optimise the performance of the baseline model. The parameters to tune are the following:

- **Length of sliding window**: 5, 10 and 20.
- **LSTM units**: 32, 64 and 128.
- **Dropout rate**: 0.2, 0.3 and 0.5
- **Learning rate**: 10^{-4} to 0.01

After correctly training the LSTM models on the dataset presented above, the baseline utility will be evaluated. Although a lot of articles only measure accuracy [10], it is important to use standard metrics for imbalanced datasets, such as AUC or F1-Score. Then, the baseline privacy will be measured using the Membership Inference Attack developed by the TensorFlow team [5], which outputs a privacy score.

With the baseline model trained, $(\epsilon, 0)$-differential privacy will be applied to the dataset, and both the utility and privacy will be measured. And with a fixed value of ϵ, the

dataset will be shrinked to investigate the impact of the dataset size in the privacy of the model using the privacy scores offered in TensorFlow.

Additionally, different values of δ will be used to apply (ϵ, δ)-differential privacy to the original dataset and, again, the model will be trained on these. Then, to assess the sensitivity of the parameter δ on the privacy of the model, both the utility and privacy will be quantified.

In order to compare the metrics appropiately, a comprehensive evaluation method will be used to analyse the trade-offs between utility and privacy. The steps will include the following components:

1. For each experiment, a performance curve will be generated to visualise how incremental changes in the privacy parameter δ or the dataset size, with a fixed ϵ value, impact the utility and privacy metrics. These curves will help identify points where the decay of those metrics is highest or lowest.

2. To validate the differences in performance among models with varying privacy parameters, hypothestis testing such as ANOVA will be conducted. This will help determine whether the observed changes are statistically significant and not due to random variation.

In addition, to determine a value for ϵ, we will use the range provided by Hsu et al. (2014) [16] for our dataset. The record space is hard to compute because the attributes in our dataset are continuous, so we can assume the worst case scenario where $|X| = N$, giving us

$$3.51 \cdot 10^{-6} \leq \epsilon \leq 10.26$$

And for δ,

$$0 < \delta \ll 3.51 \cdot 10^{-6}$$

Finally, these results will also be tested on a traditional ML model such as SVM. And the results will be reported with clear visualisations. These include performance plots, illustrating the variation of AUC, F1-Score and privacy metrics against different values of δ and dataset size, and error bars and confidence intervals to represent the variability and reliability of the results. In addition, comparisons between the ML and DL model will be studied and displayed both visually and in tables.

1.5 Work Plan and Timeline

The work plan divided into packages and the time expected for each task is described below. The duration of the projects is expected to be 13 weeks.

1.5.1 Work Packages

WP1: Literature Review and Proposal Update (expected duration: 5 days)

- Review existing literature on DP on DL models for Fraud Detection.
- Update the proposal with the help of the supervisor.

WP2: Data Acquisition, Preprocessing and Baseline Model Development (expected duration: 10 days)

- Acquire the European Credit Card Fraud Detection dataset.
- Perform data processing: normalisation and sampling.
- Train an LSTM neural network with hyperparamter tuning on the dataset and evaluate the utility to ensure good baseline performance.

WP3: Differential Privacy Varying the Dataset Size (expected duration: 17 days)

- Apply (ϵ, δ)-Differential Privacy with a fixed value of ϵ and different sizes to create multiple versions of the training data.
- Integrate these datasets into the model training pipeline and train LSTM models.
- Apply Membership Inference Attack on these models.
- Evaluate the utility and privacy metrics and compare against each other and the baseline model.

WP4: Differential Privacy Varying δ (expected duration: 12 days)

- Apply (ϵ, δ)-Differential Privacy with a fixed value of ϵ and different values of δ to create multiple versions of the training data.
- Integrate these datasets into the model training pipeline and train LSTM models.

- Apply Membership Inference Attack on these models.
- Evaluate the utility and privacy metrics and compare against each other and the baseline model.

WP5: Sensitivy Analysis and Comparison with ML Model (expected duration: 15 days)

- Train an SVM model using the same pipeline and extract the metrics.
- Generate performance curves and tables to visualise the influence of different privacy parameters.
- Conduct hypothesis testing (ANOVA) to validate the significance of the observed differences.

WP6: Reporting

- The report of the thesis will be written in parallel during the project.

1.5.2 Gantt Chart

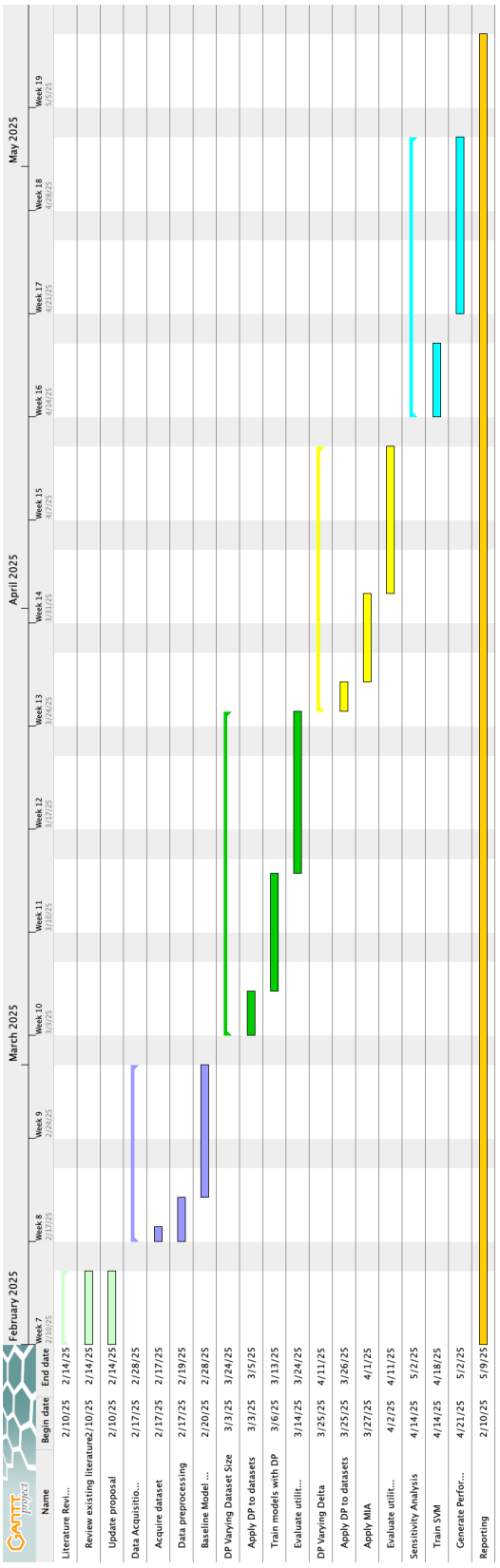

FIGURE 1.2: Gantt Chart

1.6 Ethical Issues

The most important ethical issue in this project concerns the bias and fairness of the model. As Shaham et al. [20] mention in their holistic survey, privacy and fairness are two crucial pillars in AI and ML. In this case, to prevent algorithmic bias in the model, the data needs to accurately treat all groups and individuals similarly, which can be done by making sure no personal details are taken into account to train or make predictions through differential privacy. Also, to avoid discrimination, a continuous evaluation of the predictions by the client would be needed, because they can see the original content of the dataset.

Although this project aims to enhance privacy, there is still a need to continuously making sure that the data handling methods are protecting the sensitive information properly and that there are no security risks.

Bibliography

[1] C. Dwork, "Differential privacy," in *International colloquium on automata, languages, and programming*. Springer, 2006, pp. 1–12.

[2] A. De, "Lower bounds in differential privacy," in *Theory of Cryptography: 9th Theory of Cryptography Conference, TCC 2012, Taormina, Sicily, Italy, March 19-21, 2012. Proceedings 9*. Springer, 2012, pp. 321–338.

[3] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE symposium on security and privacy (SP)*. IEEE, 2017, pp. 3–18.

[4] A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, and M. Backes, "Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models," *arXiv preprint arXiv:1806.01246*, 2018.

[5] TensorFlow, "Tensorflow privacy," https://github.com/tensorflow/privacy, 2025.

[6] S. Maes, K. Tuyls, B. Vanschoenwinkel, and B. Manderick, "Credit card fraud detection using bayesian and neural networks," in *Proceedings of the 1st international naiso congress on neuro fuzzy technologies*, vol. 261, 2002, p. 270.

[7] K. Randhawa, C. K. Loo, M. Seera, C. P. Lim, and A. K. Nandi, "Credit card fraud detection using adaboost and majority voting," *IEEE access*, vol. 6, pp. 14 277–14 284, 2018.

[8] P. Raghavan and N. El Gayar, "Fraud detection using machine learning and deep learning," in *2019 international conference on computational intelligence and knowledge economy (ICCIKE)*. IEEE, 2019, pp. 334–339.

[9] I. D. Mienye and Y. Sun, "A deep learning ensemble with data resampling for credit card fraud detection," *IEEE Access*, vol. 11, pp. 30 628–30 638, 2023.

[10] E. A. L. M. Btoush, X. Zhou, R. Gururajan, K. C. Chan, R. Genrich, and P. Sankaran, "A systematic review of literature on credit card cyber fraud detection using machine and deep learning," *PeerJ Computer Science*, vol. 9, p. e1278, 2023.

[11] L. Xiaopeng, W. Siyuan, C. Haolong, and L. Zongwei, "The utility impact of differential privacy on credit card data in machine learning algorithms," *Procedia Computer Science*, vol. 221, pp. 664–672, 2023.

[12] N. Ruan, Z. Wei, and J. Liu, "Cooperative fraud detection model with privacy-preserving in real cdr datasets," *IEEE Access*, vol. 7, pp. 115 261–115 272, 2019.

[13] J. Cai, X. Liu, and Y. Wu, "Svm learning for default prediction of credit card under differential privacy," in *Proceedings of the 2020 Workshop on Privacy-Preserving Machine Learning in Practice*, 2020, pp. 51–53.

[14] P. Basu, T. S. Roy, R. Naidu, and Z. Muftuoglu, "Privacy enabled financial text classification using differential privacy and federated learning," *arXiv preprint arXiv:2110.01643*, 2021.

[15] I. Perez, J. Wong, P. Skalski, S. Burrell, R. Mortier, D. McAuley, and D. Sutton, "Locally differentially private embedding models in distributed fraud prevention systems," in *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2023, pp. 475–484.

[16] J. Hsu, M. Gaboardi, A. Haeberlen, S. Khanna, A. Narayan, B. C. Pierce, and A. Roth, "Differential privacy: An economic method for choosing epsilon," in *2014 IEEE 27th Computer Security Foundations Symposium*. IEEE, 2014, pp. 398–410.

[17] I. D. Mienye and N. Jere, "Deep learning for credit card fraud detection: A review of algorithms, challenges, and solutions," *IEEE Access*, 2024.

[18] U. M. L. Group, "Credit card fraud detection [data set]," https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud, n.d., accessed: 2025-02-06.

[19] M. Alamri and M. Ykhlef, "Survey of credit card anomaly and fraud detection using sampling techniques," *Electronics*, vol. 11, no. 23, p. 4003, 2022.

[20] S. Shaham, A. Hajisafi, M. K. Quan, D. C. Nguyen, B. Krishnamachari, C. Peris, G. Ghinita, C. Shahabi, and P. N. Pathirana, "Holistic survey of privacy and fairness in machine learning," *arXiv preprint arXiv:2307.15838*, 2023.