Tilburg University

A COMPARATIVE ANALYSIS OF AUTOENCODERS FOR CREDIT CARD FRAUD DETECTION

ELA GUVEN

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

2067608

COMMITTEE

drs. Stijn Rotman
drs. Bosong Ding

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

June 24th, 2024

WORD COUNT

8799

ACKNOWLEDGMENTS

I would like to sincerely thank my supervisor, drs. Stijn Rotman for his guidance and support during the time I was working on my thesis. With his great help, I made a piece of work that I'm very proud of. Additionally, I would like to thank drs. Bosong Ding for taking the time to read and give feedback on this thesis.

A COMPARATIVE ANALYSIS OF AUTOENCODERS FOR CREDIT CARD FRAUD DETECTION

ELA GUVEN

Abstract

The rise in daily credit card usage has led to an increase in fraudulent transactions, resulting in significant financial losses to institutions and the government. This study investigated the effectiveness of various autoencoder models combined with a Multilayer Perceptron (MLP) classifier for credit card fraud detection using the European Cardholders dataset from Kaggle, which exhibits significant class imbalance. The Borderline-SMOTE technique was applied to address this imbalance. Three models were compared: a standalone MLP, an MLP combined with a standard autoencoder, and an MLP integrated with a Variational Autoencoder (VAE). The standalone MLP model achieved the best performance with a precision of 0.77, a recall of 0.82, and an F1 score of 0.79. The integration of a standard autoencoder with the MLP did not significantly improve performance by showing lower precision and recall. Conversely, the VAE-integrated MLP model exhibited significant enhancement over the standard autoencoder model by capturing complex data patterns and achieving a precision of 0.82 and recall of 0.66. Further analysis found that both autoencoder and VAE-integrated models exhibited high recall but low precision across different transaction amounts and temporal intervals. These findings highlight the necessity of incorporating additional features to determine fraudulent transactions accurately. The results of this thesis highlight the potential of the VAE-integrated model in capturing meaningful representations from the dataset; future research is therefore encouraged to leverage the strengths of both VAEs and MLPs on hybrid models for fraud detection systems.

1 DATA SOURCE, ETHICS, CODE, AND TECHNOLOGY STATEMENT

The research conducted in this study did not involve the collection of data from human participants or animals. The European Cardholders dataset utilized in this thesis was retrieved from Kaggle and is subject to a Database Contents License (DbCL). The data employed in this study is publicly available here for all interested parties. The author of this thesis acknowledges that they do not have any legal claim to the data. The code associated with this thesis can be accessed via the provided GitHub repository link. Part of the code was adapted from this GitHub repository and the Kaggle repository. All figures and tables presented in this study were made by the author. A generative language model, namely ChatGPT, was utilized to refine the author's original content and served as a debugging tool for resolving coding errors. For spell-checking and grammar corrections, Grammarly was used. To enhance the specificity of the writing, the author utilized Quillbot. A comprehensive overview of all software used in this study is provided in Table 7 found in the Appendix C.

2 INTRODUCTION

Anomaly detection, or outlier detection, is the process of classifying and identifying data points or patterns that are unusual or deviate significantly from the expected behavior within a dataset (Chandola et al., 2009; Injadat et al., 2018; Mehrotra et al., 2017). The primary objective of this process is to detect anomalous or abnormal data that potentially indicate errors, fraud, or suspicious activities within a given dataset. Anomaly detection is widely applicable in various domains including cybersecurity, finance, manufacturing, healthcare, and others. The significance of this task lies in its capacity to discover hidden insights, detect potential threats, or discern unexpected behavior through the examination of both normal and anomalous patterns in the dataset. Anomalies in the dataset provide crucial insights into the characteristics of the data point. For instance, unusual network activity, anomalous behavior on the heart rhythm diagram, or unauthorized credit card transactions offer critical information about unprotected data that has direct implications for individuals and society. Given the importance of these applications, the task of anomaly detection has consistently attracted attention, leading to the introduction and development of numerous approaches over the years (Habeeb et al., 2019; Mrozek et al., 2020). Furthermore, with the continuous expansion of data volume and complexity across several domains, the relevance of anomaly detection techniques will only increase (Pang et al., 2021).

Anomaly detection is extensively employed in the financial industry to detect fraud, identify market irregularities, and monitor trading activities. In the financial sector, the detection of fraudulent activities is often determined by unusual credit card transactions. In recent years, the advancement in technology has led to a rise in cashless transactions and credit card usage, which has consequently led to a notable increase in fraudulent transactions (Ounacer et al., 2018; Varmedja et al., 2019; H. Zhu et al., 2020). Given the substantial prevalence of credit card fraud, particularly in online marketplaces, the establishment of robust mechanisms for early fraud detection is imperative (Porwal & Mukund, 2018). Such mechanisms are designed to protect both customers and organizations from the adverse impacts of fraudulent transactions with high risks and costs.

Credit card fraud detection covers not only technical aspects but also represents a societal concern with significant implications for individuals and businesses on a global scale (Caroline Cynthia & Thomas George, 2021). Additionally, the effects of credit card fraud extend beyond financial losses, as it undermines trust in online transactions and can cause harm to the reputation of businesses. As a result, organizations have prioritized the development of robust fraud detection systems to maintain the integrity of financial systems.

Several recent studies have used machine learning and data mining to identify fraudulent transactions (Bin Sulaiman et al., 2022; Mrozek et al., 2020; Tiwari et al., 2021). Supervised methods like random forest, K-nearest neighborhood (KNN), support vector machine (SVM), and logistic regression are commonly used for fraud detection on a large scale compared to unsupervised methods (Bin Sulaiman et al., 2022; Ngai et al., 2011). Employing supervised methods requires a labelled dataset to classify transactions as legitimate or fraudulent. Obtaining labelled datasets in real-world situations is challenging due to computational expenses. Therefore, the effectiveness of these techniques heavily relies on training with labelled data (Ounacer et al., 2018).

On the other hand, unsupervised learning doesn't require labelling transactions as genuine or fraudulent, resulting in substantial cost savings in time and resources. As emphasized in the research by Malini and Pushpa, employing unsupervised methods for credit card fraud detection offers efficiency and flexibility. Nevertheless, these methods are not prevalent in credit card fraud detection compared to other research areas (Fanai & Abbasimehr, 2023). Hence, deep learning models and neural networks in an unsupervised manner continue to gain interest and are recognized as effective for fraud detection (Georgieva et al., 2019). Although deep anomaly detection methods have indeed made notable advancements in enhancing the effectiveness of fraud detection (Dubey et al., 2020; Majhi et al., 2019; Patidar, Sharma, et al., 2011), the complexity of this task persists due to the diverse nature of fraudulent activities (Bin Sulaiman et al., 2022; Sehrawat & Singh, 2023). Fraud detection is increasingly challenging due to the continuous evolution of fraudulent behavior patterns. Since fraudsters always find a way to show fraudulent activities as legitimate ones, two transactions may exhibit comparable attributes. Furthermore, there has been a decline in the percentage of fraudulent data points in the overall dataset as the quantity of online transaction datasets has grown. Consequently, it becomes more difficult to locate them over time (Porwal & Mukund, 2018). Due to the ever-changing nature of fraudsters' activities and their continuous development in behavior, it remains necessary to develop new approaches for credit card fraud detection (Hejazi & Singh, 2013; Malini & Pushpa, 2017; Mrozek et al., 2020). There are many ways available to improve the fraud detector.

This study employs an anomaly detection method using two types of autoencoders combined with a Multilayer Perceptron (MLP) model to enhance existing credit card fraud detectors and evaluate their effectiveness. Although limited research exists on using autoencoders for credit card fraud detection, they offer a promising approach due to their ability to gain better representations of input data (Fanai & Abbasimehr, 2023; Misra et al., 2020). Subsequently, these representations are fed into an MLP, which is used in a supervised approach to further improve fraud detection. Additionally, the autoencoders' ability to uncover hidden patterns within lower dimensional datasets makes them particularly well-suited for identifying irregularities and anomalies that may indicate fraudulent transactions (Z. Chen et al., 2018; Fanai & Abbasimehr, 2023; Pumsirirat & Liu, 2018).

3 MOTIVATION & RESEARCH QUESTIONS

Recent research on fraud detection has made significant progress by utilizing supervised and unsupervised methods. Thus, this task is often regarded as a resolved issue. Some studies have shown a high level of accuracy, exceeding 0.90 on the Credit Card Fraud Detection dataset (Dal Pozzolo et al., 2015; Ogwueleka, 2011). However, detecting credit card fraud remains difficult. A key challenge arises from the unclear nature of fraud detection datasets, where precise estimations regarding the types and prevalence of fraudulent activities remain elusive (Zamini & Montazer, 2018). In addition, machine learning models may have difficulty identifying fraudulent activities as fraud if new fraudulent transactions don't exhibit similar traits to historical data (Pumsirirat & Liu, 2018). These incidents have highlighted credit card fraud as a persistent concern that requires attention to develop more effective fraud detection systems.

Notwithstanding progress in current methodologies, there are still limitations in the following areas concerning credit card fraud detection:

- **The absence of accessible public datasets for credit card fraud detection**: Banks and financial institutions are now increasingly able to obtain large-scale credit card transaction datasets due to the growing popularity of e-payment methods. Unfortunately, having access to these credit card transaction datasets for investigation purposes is highly challenging due to security and data privacy concerns, as highlighted in the study by H. Zhu et al. Banks are hesitant to disclose their customers' data explicitly because of the General Data Protection Regulation (GDPR), as discussed in a review by Bin Sulaiman et al. The combination of these factors makes it hard to find a public dataset, thereby hindering the development of adequately trained fraud detection systems (Saia & Carta, 2019).

  The European Cardholders dataset, as introduced by Dal Pozzolo et al., is widely used for training fraud detection algorithms due to its incorporation of daily credit transactions. While this dataset might encourage progress in credit fraud detection, a substantial training dataset is still necessary to comprehensively address fraudulent activities (Bin Sulaiman et al., 2022). Training the algorithms on new datasets will significantly enhance feature extraction from the dataset and improve fraud detection dramatically. Therefore, credit card fraud detection can be improved by analyzing additional datasets containing a variety of fraudulent activities.

  Collecting a large number of fraudulent transactions in a single dataset poses a significant challenge, and it may not be feasible to include all different types of fraud (Mrozek et al., 2020; Tingfei et al., 2020). For instance, certain public datasets, like the German dataset (Patil et al., 2018), and private banks' datasets exhibit a lack of diversity in fraudulent activities. The efficacy of fraud detectors relies heavily on the input data (Mrozek et al., 2020). Hence, this limitation presents a significant obstacle to developing robust and accurate fraud detection algorithms that can effectively handle real-world situations (Kim et al., 2019).

- **High false-negative rate**: Models on credit card fraud detection are usually trained on the widely used European Cardholders dataset (Dal Pozzolo et al., 2015) and assessed based on the dataset's validation set. Typically, results in the studies are compared based on the accuracy (Asha & KR, 2021; Carcillo et al., 2018; Razooqi et al., 2016). Instead of relying solely on accuracy, it is better to use other evaluation metrics such as precision, recall, F1 score, or the area under the receiver operating characteristic curve (AUC- ROC) to offer a more comprehensive assessment of the model on unseen data.

The review conducted by Makki et al. highlights that relying on a single performance metric can result in misinterpretation, especially in imbalanced datasets. As a result, achieving better outcomes on other metrics will still improve this task.

Fraud detection in the financial sector remains challenging even with the implementation of neural networks (X. Zhu et al., 2021). This difficulty stems from the difficulty of evaluating the changing techniques employed by fraudsters over time. The recent fraud detection algorithms have high accuracy, but the precision and recall of these models are typically low (Dal Pozzolo et al., 2015; Mrozek et al., 2020). In this scenario, fraud detection models often mismatch fraudulent transactions with legitimate ones. Naturally, this raises the question of the model's ability to detect fraudulent observations. Consequently, models that minimize the number of false negatives during the detection process are critical to improve performance and prevent financial losses for banks.

- **Class imbalance problem**: Anomaly detection tasks often encounter a significant challenge because of the imbalanced nature of the datasets. An unbalanced dataset in credit card fraud detection tasks refers to a situation where the number of legitimate (non-fraudulent) transactions far exceeds fraudulent transactions. The presence of an imbalance in the dataset can lead to biased models, higher false-negative rates, and difficulty accurately evaluating the performance of fraud detectors. As a prominent obstacle in credit card fraud detection, the class imbalance issue is extensively discussed in the literature (Makki et al., 2019). This problem is characterized by a highly imbalanced and skewed data distribution (Hilal et al., 2022; Makki, 2019). This issue has led to significant financial losses as a consequence of the inability of machine learning models to accurately identify fraudulent activities. Various methods have been suggested in the literature to address class imbalance (Fernández et al., 2018; Haixiang et al., 2017), but they are still not adequately effective for fraud detection (Johnson & Khoshgoftaar, 2019; Makki et al., 2019). Consequently, addressing this problem is necessary to enhance the performance of fraud detection systems.

  The approach to addressing this issue depends on the degree of imbalance in the dataset; therefore, there are no specific balancing techniques suited for individual models (Makki et al., 2019). One major drawback of balancing approaches is that they yield varying outcomes with particular types of models and datasets (Zareapoor et al., 2012).

This thesis aims to contribute to credit card fraud detection by addressing the challenges mentioned above. Effective credit card fraud detection systems must accurately identify fraudulent activities commonly occurring in daily financial transactions. To achieve this, the thesis utilizes two autoencoders combined with a Multilayer Perceptron (MLP) model to improve credit card detection by reducing false negatives. This study aims to demonstrate the effectiveness of various autoencoder architectures in detecting fraud by capturing complex relationships within the data. Furthermore, this thesis provides insights into the performance of different autoencoder models when combined with a class-balancing technique such as Borderline-SMOTE. By implementing these methods, financial institutions can better protect against potential financial losses; thereby, they can maintain customers' trust in banks by ensuring the proper functioning of the financial system and supporting overall economic activity. Institutions that fail to implement effective fraud detection measures face the possibility of rising costs and decreasing consumer spending, which can adversely affect businesses and governments. An efficient fraud detection system can mitigate these effects.

The research question in this thesis can be formulated as follows:

> *"To what extent can different types of autoencoders effectively distinguish fraudulent credit card transactions from legitimate ones?"*

The following sub-questions must be addressed in order to provide an answer to the main research question:

SQ1 *"How does the transaction amount influence the autoencoders' ability to detect fraudulent transactions?"*

This question investigates how the size of transactions throughout the day affects the ability of autoencoders to distinguish between fraudulent and legitimate transactions. By analyzing how transaction sizes throughout the day affect the models' performance, it aims to uncover any correlations between transaction amounts and the effectiveness of fraud detection models.

SQ2 *"How does the time difference between transactions affect the performance of autoencoders in detecting fraud?"*

This question delves into the temporal aspect of autoencoder performance in fraud detection. Specifically, it examines the models' ability to identify fraudulent activities in consecutive transactions and explores any relationships between fraudulent behavior and time intervals.

4 RELATED WORK

4.1 *Credit card fraud detection using autoencoders*

Autoencoders, as feed-forward multilayer neural networks, have recently gained considerable interest in many research fields for their capacity to uncover hidden patterns within input data (Bengio et al., 2012; Wang et al., 2015).These neural networks receive high-dimensional input data, compress it into a lower-dimensional latent space, and then reconstruct data in the output layer. By compressing the data into a representation with fewer dimensions, the autoencoders can extract meaningful features from complex patterns within the data, resulting in efficient storage and faster computations for algorithms (Fournier & Aloise, 2019; Liu et al., 2017).

Autoencoders have emerged as valuable tools for anomaly detection, including credit card fraud detection (Chaquet-Ulldemolins et al., 2022; Chow et al., 2020; Lin & Jiang, 2020; Rezapour, 2019; Zhou & Paffenroth, 2017). Unlike Principal Component Analysis (PCA), which may not capture non-linear feature correlations effectively, autoencoders are more effective in finding anomalies by capturing non-linear relationships (Z. Chen et al., 2018; Hinton & Salakhutdinov, 2006; Niu et al., 2019; Sakurada & Yairi, 2014). Studies by Schreyer et al. and Zheng et al. have shown that autoencoders achieve lower rates of false negatives compared to traditional rule-based systems.

Prior research in fraud detection has demonstrated that autoencoders offer computational efficiency and performance enhancements, surpassing traditional machine learning approaches (Paula et al., 2016; Renström & Holmsten, 2018). Despite their advantages, autoencoders have not been widely employed in credit card fraud detection compared to other machine learning techniques (Roy et al., 2018; Xuan et al., 2018). Existing research indicates that training autoencoders on larger, real-time datasets can enhance fraud detection, although the lack of publicly accessible datasets limits progress (Kazemi & Zarrabi, 2017; Pumsirirat & Liu, 2018; Singla et al., 2020). Hence, it is imperative to explore the capabilities of autoencoders to improve fraud detection in this specific domain.

Recent studies have combined autoencoders with classification models, using autoencoders to extract essential data structures while filtering out noise (Al-Shabi, 2019; Lin & Jiang, 2021; J. Zou et al., 2019). Furthermore, some studies have shown that this combination often improves performance, particularly with larger training datasets (Ouedraogo et al., 2021; Pumsirirat & Liu, 2018; Zioviris et al., 2022).

However, there remains a significant research gap regarding the extensive use of autoencoders for credit card fraud detection. In the proposed approaches, autoencoders are typically trained on non-fraud transactions (Al-Shabi, 2019; Chalapathy et al., 2018; Chaquet-Ulldemolins et al., 2022; Zamini & Montazer, 2018; Zhou & Paffenroth, 2017). Consequently, contemporary academic work has increasingly focused on techniques involving the training of autoencoders utilizing data samples that do not contain anomalies (Du et al., 2023). Therefore, their potential may be somehow restricted (Zioviris et al., 2022).

Furthermore, there is limited research on the comparative analysis of various types of autoencoders within this research domain. Previous studies by Renström and Holmsten, A. Ali et al., Tingfei et al. and Kumar have suggested that both traditional and variational autoencoders (VAEs) can enhance the precision and accuracy of credit card fraud detection. Some research indicates superior performance of VAEs in the European Cardholders dataset (Ouedraogo et al., 2021; Raza & Qayyum, 2019; Sweers et al., 2018), while others find standard autoencoders to be more effective for fraud detection (Pumsirirat & Liu, 2018; Zioviris et al., 2022).

In light of these findings, this thesis aims to fill the gap by conducting a comparative analysis of different autoencoder architectures combined with an MLP model for credit card fraud detection.

4.2 *Techniques for handling imbalanced data in credit card fraud detection*

Detecting credit card fraud presents a formidable challenge characterized by data imbalance, where the number of legitimate transactions significantly exceeds the fraudulent instances (Shamsudin et al., 2020). Managing this class imbalance problem is crucial for developing accurate and reliable fraud detection systems (Singh et al., 2022). Unbalanced datasets negatively impact model performance and reduce the effectiveness of various classification methods (C. P. Chen & Zhang, 2014; Lucas & Jurgovsky, 2020; Shamsudin et al., 2020). Therefore, strategies for addressing this issue play a critical role for models in enhancing the fraud detection rate.

Numerous methodologies, including oversampling, undersampling, and cost-sensitive learning, have been proposed to address data imbalances in credit card fraud detection (López et al., 2013; Makki et al., 2019; Rout et al., 2018; Vandewiele et al., 2021). Unlike algorithmic modifications, sampling-based techniques have gained significant popularity due to their simplicity and computational efficiency (Alamri & Ykhlef, 2024; Ebenuwa et al., 2019). By adjusting the distribution of the training data through oversampling or undersampling, these techniques aim to alleviate the skewness towards the majority class, thus enabling classifiers to generalize better and make accurate predictions (Abd El-Naby et al., 2023; H. Ali et al., 2019; Tyagi & Mittal, 2020). Moreover, sampling methods can be easily integrated into existing machine learning pipelines without significant limitations (Alam et al., 2022).

According to Sisodia et al., the first step to acquire accurate findings from algorithms is to prepare training data at the preprocessing stage. Hence, selecting appropriate sampling methods for imbalanced datasets becomes imperative due to their substantial impact on the performance of detection models (Mrozek et al., 2020). While oversampling methods involve generating synthetic instances of minority fraudulent classes to balance the class distribution, undersampling aims to reduce the majority of legitimate class samples to achieve a balanced dataset.

Oversampling techniques, such as the Synthetic Minority Oversampling Technique (SMOTE), Adaptive Synthetic Sampling (ADASYN), and random oversampling, have been effectively employed with different classifiers across numerous datasets (Aguiar et al., 2023). Previous research shows that oversampling techniques outperform undersampling methods, which may lead to the loss of valuable information from the majority class (Ahmad et al., 2023; Ebenuwa et al., 2019; Lunghi et al., 2023).

Among these techniques, SMOTE is particularly popular for generating a balanced dataset in credit card fraud detection (Almhaithawi et al., 2020; Ileberi et al., 2021; Prasetiyo et al., 2021). However, SMOTE has limitations, such as the possibility of synthetic data overlapping with majority class samples or generating irrelevant synthetic instances where the decision boundary is ambiguous (Ebenuwa et al., 2019; Shamsudin et al., 2020; Strelcenia & Prakoonwit, 2023a). As a result, Borderline-SMOTE was developed, focusing on synthesizing examples near the decision boundary, where classification is more challenging (Alamri & Ykhlef, 2022; De La Bourdonnaye & Daniel, 2022). Since the class boundary between legitimate and fraudulent transactions can be intricate and dynamic, Borderline-SMOTE has emerged as a valuable oversampling technique in credit card fraud detection by enhancing the discriminatory power of classifiers and improving fraud detection accuracy (Le, 2022; Patra et al., 2023; Xie et al., 2023).

In addition to sampling techniques, Generative Adversarial Networks (GANs) have been explored for generating new data samples to tackle class imbalance, with some studies demonstrating their superiority over other oversampling methods (Fiore et al., 2019; Strelcenia & Prakoonwit, 2023b; Tingfei et al., 2020). However, GANs have limitations, including a lack of diversity within synthesized instances and the need for larger datasets, which may pose practical challenges(Hung & Gan, 2021). Therefore, implementing and training GANs require additional adjustments and computational resources (Aftabi et al., 2023; Gangwar & Ravi, 2019; Ghaleb et al., 2023; Pandey et al., 2020).

This study utilizes Borderline-SMOTE as the balancing technique to enhance the effectiveness of the fraud detection model. Borderline-SMOTE offers a simple method of generating synthetic instances by facilitating easy implementation with different models. Additionally, its proven effectiveness in improving fraud detection performance and addressing class imbalances without significant computational overhead makes it a practical and reliable choice over GANs for balancing imbalanced datasets in credit card fraud detection (Alamri & Ykhlef, 2022; Obimbo et al., 2021; Sun et al., 2022; H. Zou, 2021).

5 METHOD

This section outlines the methodology and experimental framework employed in this study. Firstly, a description of the European Cardholders dataset is provided. The subsequent analysis and preprocessing of the data are explained. The primary approach undertaken in this research involves a comparative analysis of two types of autoencoder models for detecting credit card fraud. The general structure of autoencoder models is explained in order to clarify the functioning of these models. The algorithms used in this thesis are (1) Standard Autoencoders, (2) Variational Autoencoders (VAEs), and (3) Multilayer Perceptron (MLP) as classifier. Finally, evaluation metrics used in the performance assessment are specified. The methodology's workflow is depicted in Figure 1.

5.1 *Dataset Description*

Due to the lack of publicly available datasets, the proposed method was evaluated using a dataset obtained from European Cardholders (Dal Pozzolo et al., 2015). This dataset contains a compilation of credit card transactions conducted by individuals during two days in September 2013. The dataset consists of 284,807 transactions, with 492 instances flagged as fraudulent, accounting for 0.172% of the total transactions. This dataset is highly imbalanced, with a significantly smaller proportion of fraudulent activities in comparison to legitimate transactions. Consequently, this problem must be handled carefully to build an effective fraud detector.

The dataset comprises important numerical input attributes that have been acquired through PCA transformation. This process has been deployed to retain meaningful features while reducing dimensionality to protect user identities and confidential information. Although these features are derived from transaction records, their contextual details are

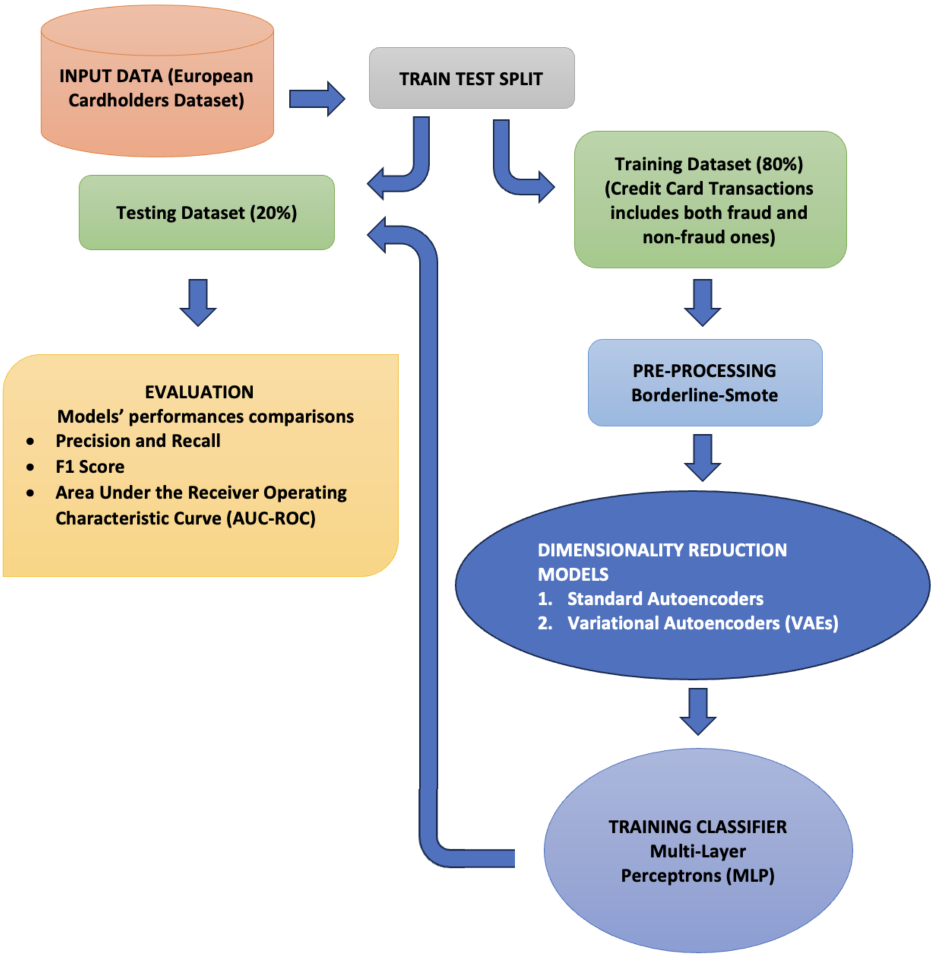
INPUT DATA (European Cardholders Dataset)
TRAIN TEST SPLIT
Testing Dataset (20%)
Training Dataset (80%) (Credit Card Transactions includes both fraud and non-fraud ones)
PRE-PROCESSING Borderline-Smote
EVALUATION
Models' performances comparisons
Precision and Recall
F1 Score
Area Under the Receiver Operating Characteristic Curve (AUC-ROC)
DIMENSIONALITY REDUCTION MODELS
1. Standard Autoencoders
2. Variational Autoencoders (VAEs)
TRAINING CLASSIFIER Multi-Layer Perceptrons (MLP)

Figure 1: Overview Methodology

anonymized to address privacy concerns. Moreover, the dataset incorporates attributes such as "time," "class," and "amount." The "amount" feature denotes the transaction amount, whereas the "time" attribute indicates the duration in seconds since the initial transaction. Lastly, the "class" variable distinguishes between fraudulent transactions (denoted by 1) and non-fraudulent transactions (denoted by 0).

5.2 *Exploratory Data Analysis and Preprocessing*

Exploratory Data Analysis (EDA) provides valuable insights into the characteristics of a dataset and visualizes the relationships within the transaction data. Therefore, several types of EDA have been implemented to better understand the data features in the dataset. The insights gained from this analysis are critical for the subsequent modeling phase, where they guided model tuning to improve fraud detection capabilities.

- **Missing Value Analysis:** The dataset has been verified for missing values and is found to be complete.

- **Distribution Analysis:** Histograms and KDE (Kernel Density Estimation) plots visualize the distributions of transaction amounts and times for both fraudulent and non-fraudulent transactions. Figure 20 in Appendix B shows that non-fraudulent transactions occur more frequently and maintain a consistent frequency over the observed time period. Conversely, fraudulent transactions are relatively sparse and exhibit less variability over time. Figures 21 and 22 in Appendix B illustrate that fraudulent transactions involve a broader range of transaction amounts and tend to be smaller than non-fraudulent transactions.

  Additionally, scatter plots of transaction time versus amount for each class are presented in Appendix B. Non-fraudulent transactions do not display a clear correlation between time and amount, whereas fraudulent transactions are more dispersed across both variables, suggesting potential time-dependent patterns in fraudulent activity, which could be crucial for anomaly detection.

- **Dimensionality reduction and visualization:** Dimensionality reduction techniques, such as t-SNE, PCA, and Truncated SVD (tSVD), are applied to explore the relationships among features and their association with the target variable, which is the transaction class. These techniques help visualize the relationships of variables in a smaller sample, which closely resembles the compressed representation of input data by autoencoders. In this dataset, most features were already transformed using PCA to retain meaningful features while reducing dimensionality for privacy reasons. The additional dimensionality reduction techniques employed in this step further analyze the transformed features.

  The t-SNE plot, referenced in Figure 24, suggests a complex, nonlinear relationship between the two classes. Furthermore, Figures 25 and 26 demonstrate that the PCA and tSVD plots do not effectively separate the fraud cases from the non-fraudulent cases in this dataset. This outcome underscores the challenge of identifying clear boundaries between classes using linear dimensionality reduction techniques alone, thereby justifying the utilization of autoencoders in this study.

- **Correlation Analysis:** The correlation matrix of features, shown in Appendix B exhibits various degrees of correlation between the anonymized features and the class label. Some features demonstrate a positive or negative correlation with class, indicating their potential utility in distinguishing between fraudulent and non-fraudulent transactions. In particular, the correlation between "Time" and "Class" is approximately zero, indicating no direct linear relationship. Nevertheless, the absence of a direct relationship between variables does not exclude the possibility of complex, time-dependent patterns in fraudulent transactions that simple correlations fail to capture. To understand these time-dependent patterns, it is essential to analyze the distribution of other variables, such as "Amount," over time for both fraudulent and non-fraudulent transactions, as detailed in the Appendix.

- **Class Imbalance Analysis:** Figure 2 highlights the severe class imbalance within the dataset by showing the percentage of fraudulent and non-fraudulent transactions. This is crucial for understanding the dataset's skew towards non-fraudulent transactions and helps in planning how to handle imbalanced data during model training.

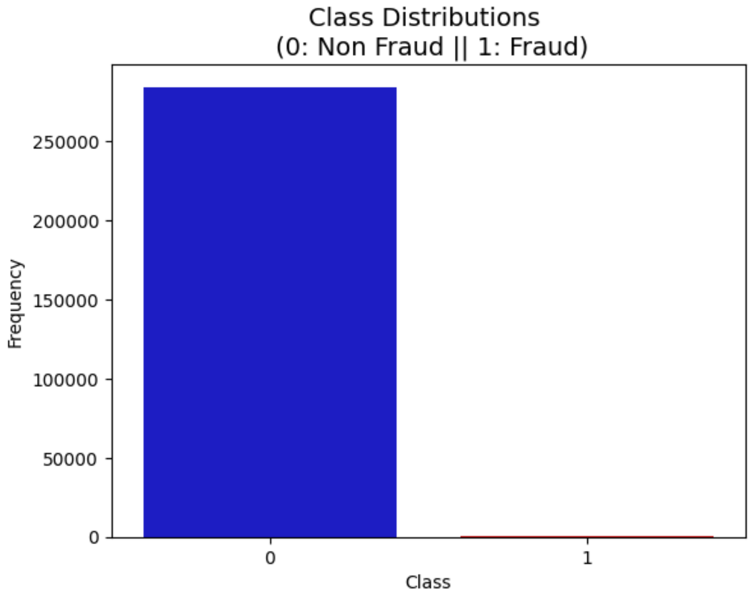

Figure 2: Class Distribution

Each of these EDA techniques offers crucial insights into the data, reveals underlying patterns, and assists in addressing questions regarding how different features may relate to the occurrence of fraud. These insights are invaluable for building effective fraud detection mechanisms.

A comprehensive exploratory data analysis (EDA) was conducted on the European Cardholders dataset, followed by a series of preprocessing steps. These steps included data normalization and the application of the oversampling technique after the data was divided into training and testing sets.

The dataset was previously standardized, with most features having a mean close to zero, except for the 'Time' and 'Amount' columns. During the preprocessing phase, the 'Time' and 'Amount' variables were scaled to a smaller range to enhance the efficiency of the training process.

The dataset was split into training and test sets with proportions of 80% and 20%, respectively. Given the infrequency of fraud cases, the Borderline-SMOTE technique was employed on the training dataset to ensure an adequate number of anomalous transactions during the training process. To preserve the integrity of real-world conditions and prevent data leakage, oversampling was confined to the training set only.

5.3 *Algorithms*

This subsection describes the algorithms utilized in this study, including their architectures and associated hyperparameters. The models employed were Autoencoders, Variational Autoencoders (VAEs), and Multilayer Perceptrons (MLP). This comprehensive description aims to provide a precise comprehension of the implementation of these algorithms in this study.

5.3.1 *Autoencoders*

In credit card fraud detection, autoencoders can serve several roles with dimensionality reduction, feature learning, and anomaly detection. An autoencoder, as a generative deep learning algorithm, is comprised of two main components: an encoder and a decoder module. The encoder module is responsible for compressing the input data into a lower dimensional representation and conversely, the decoder module reconstructs this representation to the original input data. The architecture of the autoencoders is illustrated in the figure below.

Autoencoders are designed to reproduce the same amount of output data from the input data. The core architecture of an autoencoder comprises essential components that facilitate the extraction and representation of significant features within the dataset. This approach improves computational efficiency by reducing dimensions during training. It also helps in mitigating data noise and enhances model performance by focusing

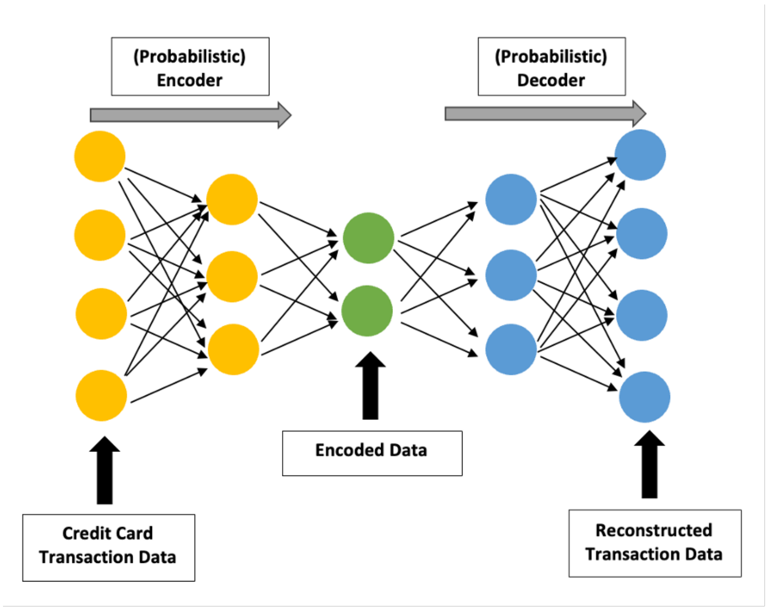

Figure 3: Visualization of Autoencoders Architecture

on relevant features. The encoding and decoding processes facilitated by autoencoders effectively capture complex patterns and anomalies within transaction data that might not be immediately apparent in raw data. Consequently, autoencoders are considered a highly promising tool for fraud detection, as they identify intricate patterns that are not captured through traditional methods (Z. Chen et al., 2018; Cheng et al., 2020). This process is achieved by the model's ability to approximate an identity function through learning the compressed representation of the input data.

$$f_{W,b}(x) \approx x$$

Where $f_{W,b}(x)$ is the reconstructed output data, x is the input data, W are the weights, and b are the biases.

In the context of credit card fraud detection, autoencoders are trained typically on non-fraudulent transactions (Sharma et al., 2022). Autoencoders reconstruct the input and optimize the parameters by minimizing the reconstruction error. Typically, mean squared error is used as a reconstruction error for flagging anomaly.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2$$

where n is the size of the data, x is the input data, and $\hat{x}$ is the reconstructed output data.

In this generic strategy, the autoencoder learns to reconstruct the normal transactions very well, and when a fraudulent transaction, which is different in nature from the 'normal' transactions, is passed through the autoencoder, it results in a higher reconstruction error. This elevated error is directly used to determine a threshold to classify transactions as fraudulent or legitimate, where a high error suggests fraudulent activity and a low error indicates a legitimate transaction.

In line with current methodologies employing autoencoders for anomaly detection, the primary objective of this thesis is to utilize autoencoders for dimensionality reduction before classification. Traditionally, fraud detection methods assume that fraudulent cases are rare and significantly different from each other (Awosika et al., 2024; Baesens et al., 2021; Hilal et al., 2022). However, recent studies have demonstrated that fraudulent transactions can share similar characteristics, like normal transactions (Benchaji et al., 2021; Davidson, 2022). Therefore, this thesis diverges from conventional literature by training autoencoders on a dataset comprising both fraudulent and non-fraudulent transactions. This approach contrasts with the typical use of autoencoders for anomaly detection, where they are often trained exclusively on 'normal' data.

5.3.2 *Variational Autoencoders (VAEs)*

Variational autoencoders (VAEs) are a variant of autoencoders that fall under the category of generative models. They integrate the architecture of autoencoders with probabilistic approaches, specifically designed to learn the distribution of input data and generate new data that resembles the training data. A Variational Autoencoder (VAE) is composed of an encoder, which transforms input data into a distribution in the latent space, and a decoder, which subsequently reconstructs the data from this latent space.

The VAE framework includes latent variables z sampled from a Gaussian distribution parametrized by the encoder as follows:

$$z \sim q(z \mid x) = \mathcal{N}(\mu(x), \sigma(x)^2)$$

where $\mu(x)$ and $\sigma(x)$ are functions of the input data x that produce the mean and standard deviation of the latent distribution, respectively.

VAEs employ a method where inputs are encoded as distributions rather than single points in order to create a regularized latent space. VAEs offer a significant benefit by incorporating probabilistic modeling into the encoding and decoding process, enhancing the generative capabilities of autoencoders (Kingma & Welling, 2013). Traditional autoencoders often struggle to generate new data due to irregularity in the latent space, potentially resulting in irrelevant data points during decoding.

Additionally, VAE optimizes a combined loss function that includes both the reconstruction loss and a regularization term. The reconstruction loss, measured by mean squared error, ensures accurate data reconstruction. The regularization term, Kullback-Leibler (KL) divergence, ensures that the encoded distributions closely approximate a standard normal distribution.

The combined objective function is given by:

$$\mathcal{L} = \mathbb{E}_{q(z|x)} [\log p(x \mid z)] - D_{\mathrm{KL}} (q(z \mid x) \parallel p(z))$$

where $\mathbb{E}_{q(z|x)} [\log p(x \mid z)]$ *represents the expected value of the distribution* $q(z \mid x)$ *and the logarithm of the conditional probability of* x *given* z*, denoted as* $\log p(x \mid z)$*. The expression* $D_{\mathrm{KL}} (q(z \mid x) \parallel p(z))$ *is the Kullback-Leibler divergence between the conditional distribution* $q(z \mid x)$ *and the prior distribution* $p(z)$*.*

Thus, the VAE Loss is a combination of Reconstruction Loss and KL Divergence:

$$\textit{VAE Loss} = \textit{Reconstruction Loss} + \textit{KL Divergence}$$

The primary goal of training the VAE model is to minimize both the reconstruction error and the KL divergence of the latent variables. The VAE model can obtain a continuous representation in the latent space using variational inference, which allows for a certain degree of interpretability. VAEs have become a popular method for modeling complex generative distributions as an extension of standard autoencoders.

VAEs have a wide range of applications, such as generative models for creating new data points, dimensionality reduction, and anomaly detection. They generate probability distributions over reconstructed data, thereby allowing the confidence interval for identifying anomalies. Several studies have explored the use of VAEs for fraud detection.

In literature, VAEs are often used as an oversampling module to address the class imbalance issue, generating new instances of minority class for training the model. For instance, Ibrahim et al.; Tingfei et al. employed VAEs to create synthetic data to balance datasets, significantly improving the performance of fraud detection models. Furthermore, Ding et al. demonstrated that VAEs could effectively capture the complex patterns of fraudulent behavior, leading to more robust detection mechanisms.

Additionally, a VAE-based anomaly detection framework was introduced for unsupervised fraud detection, where they trained on legitimate transactions and the reconstruction error was used to identify fraudulent transactions. This approach indicated that VAEs could autonomously distinguish between normal and fraudulent transactions by learning the underlying data distribution and recognizing patterns (Alazizi et al., 2020; An & Cho, 2015; Anh et al., 2020; Ouedraogo et al., 2021; Shen, 2021).

In this thesis, both traditional autoencoders and VAEs help in reducing the dimensionality of the data and learning efficient representations that might capture underlying patterns in the data. Each has its strengths, with VAEs providing a probabilistic manner for describing the data, which might capture nuances in how data varies.

5.3.3 *Multilayer Perceptron*

A Multi-Layer Perceptron (MLP) is a feedforward neural network that consists of an input layer, multiple hidden layers, and an output layer. As a supervised learning algorithm, MLP was trained using labelled data. Throughout the training process, the network learns to map input data to correct labels by iteratively adjusting the weights and biases to minimize the error.

The functional form of the MLP can be expressed as follows:

$$y = \varphi \left(\sum_{i=1}^{n} \omega_i x_i + b \right)$$

where x represents input values, w denotes the weights, b is the bias, and φ is the non-linear activation function.

MLPs have been extensively employed in fraudulent transaction detection due to their ability to learn complex relationships within transaction data. Empirical evidence suggests that MLPs outperform traditional machine models, achieving higher detection accuracy (Abd El Naby et al., 2021; Mishra & Dash, 2014; Moumeni et al., 2022; Riffi et al., 2020; Shirodkar et al., 2020).

Advancements in deep learning have further enhanced the performance of MLPs in fraud detection by incorporating deeper architectures with multiple hidden layers and advanced optimization techniques (Kasasbeh et al., 2022; Pillai et al., 2018). Additionally, recent studies demonstrated that MLPs can be integrated with other models to create a hybrid approach, and this enhances the robustness of the fraud detection systems, resulting in improved precision and recall without significantly increasing false positives for legitimate transactions (Pumsirirat & Liu, 2018).

Given that the MLP represents the simplest form of artificial neural networks (ANNs) and considering the objective of this study, which is to compare the effectiveness of two types of autoencoders, the MLP was chosen as the benchmark classifier.

5.3.4 *Hyperparameter Tuning*

Hyperparameter tuning is crucial in developing effective deep learning models by selecting the optimal set of hyperparameters. This process enhances the computational efficiency, robustness, and accuracy of the models by mitigating issues such as overfitting and underfitting. Proper hyperparameter tuning directly impacts the performance of models in fraud detection (Taha & Malebary, 2020). Each model was optimized to achieve its best possible performance to achieve a fair comparison among the three approaches. During the tuning process, considerations for faster training times and reduced computational resources were significantly emphasized.

In tuning the Multilayer Perceptron (MLP), various settings were explored. Based on the classifier results, the number of layers and the number of neurons in each layer were incrementally increased to enhance model complexity and performance. Different dropout rates were tested to determine the optimal level of regularization. Batch sizes were varied from 32 to 512 to assess their impact on training time. The hyperparameter settings for the MLP are presented in Table 1 below.

Hyperparameter	**Values**
learning_rate	0.1, 0.01, 0.001
batch_size	128, 256, 512
epochs	30, 50
activation	relu, tanh, sigmoid
dropout_rate	0.1, 0.3, 0.5
optimizer	Adam, SGD
dense_units_layers	64, 128, 256

Table 1: Hyperparameter Settings for MLP

To augment the complexity and efficacy of the autoencoder models in capturing and reconstructing intricate data patterns, hyperparameter tuning was systematically applied (Table 2 and 3). Additional layers were incorporated to increase model complexity, while advanced techniques such as dropout were used to regularize the models and reduce overfitting. Furthermore, batch normalization was implemented to promote training stability by ensuring consistent normalization across layers, and various activation functions were tested to enable the models to capture more complex data representations. The number of epochs was adjusted, and early stopping was employed to ensure convergence. These strategic modifications were crucial in optimizing the model's performance on unseen data.

Given that this thesis compares the effectiveness of two autoencoders in credit fraud detection, the same MLP architecture was tested using the encoded features as inputs from both traditional autoencoders and Variational Autoencoders (VAEs).

Hyperparameter	Values
autoencoder_learning_rate	0.1, 0.01, 0.001, 0.0001
autoencoder_epochs	50, 100
autoencoder_batch_size	32, 256, 512
encoder_units_layers	13, 6, 3
dropout_rate	0.1, 0.3, 0.5
activation	relu, tanh, sigmoid

Table 2: Hyperparameter Settings for Autoencoder + MLP

Hyperparameter	Values
vae_learning_rate	0.1, 0.01, 0.001, 0.0001
vae_epochs	50, 100
vae_batch_size	128, 256, 512
encoder_units_layers	512
latent_dim	2, 4
dropout_rate	0.1, 0.2
activation	relu, leakyrelu

Table 3: Hyperparameter Settings for VAE + MLP

5.4 *Experimental Set-up*

This section details the experiments conducted to address the research questions related to the efficacy of various autoencoder types in distinguishing fraudulent credit card transactions from legitimate ones.

Experiment 1: Multilayer Perceptron (MLP) Model with resampled credit card transaction data

Before the first experiment, the Borderline-SMOTE technique was used to address the issue of class imbalance by creating synthetic samples for the minority class near the decision boundary. This preprocessing step ensured that all subsequent models were trained on a balanced dataset. Subsequently, a Multilayer Perceptron (MLP) was implemented and evaluated on the resampled data. The objective of this experiment was to establish baseline performance and evaluate how implementing the autoencoders to the pipeline affects the ability of MLP to detect fraudulent transactions.

Experiment 2: Standard Autoencoder and MLP Integration

The second experiment involved the integration of a standard autoencoder with the MLP classifier. In this experiment, autoencoders are used for preparing data for subsequent supervised learning task. The autoencoder was trained on the resampled training data to reconstruct input data.

Additionally, the trained autoencoder reconstructed the test data, and the mean squared error (MSE) between the original and reconstructed data was calculated. The reconstruction error was plotted for both normal and fraudulent transactions, and a threshold was determined to distinguish between the two classes. An encoder model was created to generate encoded features from the input data, which are then used to transform both the training and test data. The MLP was trained on the encoded training data, and subsequently, the MLP model was evaluated on the encoded test data to determine its performance.

This experiment aimed to investigate whether the latent representations learned by the autoencoder could enhance the MLP's performance in identifying fraudulent transactions. The effectiveness of this integrated model was compared against the baseline MLP established in Experiment 1.

Experiment 3: Variational Autoencoder (VAE) and MLP Integration The third experiment introduced Variational Autoencoders (VAEs) in conjunction with the MLP classifier. VAEs impose a probabilistic structure on the latent space, potentially capturing more meaningful data representations (Z. Chen et al., 2021). As in Experiment 2, the reconstruction error from the VAE was utilized as an additional feature next to the MLP classifier.

This experiment intended to determine whether the probabilistic nature of VAEs provided any enhancements in performance compared to the standard autoencoder in detecting fraudulent transactions.

Additional Experiments for Sub-Questions To address the research sub-questions, further experiments were conducted using the same models with test data segmented based on transaction amounts and time intervals.

- ***Transaction Amount Analysis:*** The performance of autoencoder-based models (Standard Autoencoder + MLP, VAE+MLP) was evaluated across different transaction amounts to understand how transaction size influences the model's efficacy in fraud detection. The test dataset was segmented into low and high transaction amounts. Each test segment was encoded using the trained encoder model. The MLP which was trained on the encoded features from the full training dataset, was then evaluated on each test segment.

- ***Time Difference Analysis:*** The models were also evaluated based on the time difference between transactions to uncover any relationships between transaction timing and the models' performance in fraud detection. The test set was segmented into short and long intervals based on the time differences between consecutive transactions.

Each test segment was encoded using the trained encoder model. The MLP which was trained on the encoded features from the full training dataset, was then evaluated on each segment. Finally, the performance of the MLP classifier for each segment was analyzed to determine how temporal differences impact the performance of the autoencoders-MLP pipeline.

5.5 *Evaluation Metrics*

The primary objective of this thesis is to evaluate and compare the performance of two types of autoencoders integrated with Multilayer Perceptron (MLP) for credit card fraud detection, utilizing the European Cardholders dataset. This study aims to determine which model exhibits superior performance in terms of reconstruction error and other relevant evaluation metrics. This research strives to enhance credit card fraud detection, enabling the accurate identification of fraudulent transactions while minimizing the occurrence of false negatives.The evaluation of the credit card fraud detection models was conducted using a confusion matrix.

Integrating autoencoders and Variational Autoencoders (VAEs) into a fraud detection pipeline allows the MLP to benefit from the more discriminative and compact representations learned by autoencoders. This integration aims to improve the model's ability to classify fraudulent transactions, thereby enhancing overall fraud detection capabilities.

The effectiveness of the autoencoder models was assessed based on metrics such as reconstruction error and loss function. Reconstruction error quantifies the difference between the input data and the reconstructed output from the autoencoder, with lower reconstruction errors indicating better performance. The loss function used during training, such as mean squared error (MSE), optimizes the model by minimizing the difference between the predicted and actual values. For the Multilayer Perceptron (MLP) classifier, performance was evaluated using precision, recall, F1 score, and the Area Under the Receiver Operating Characteristic Curve (AUC score).

6 RESULTS

This section aims to present and explain the results of experiments mentioned in Section 5.4.

6.1 *Results of Experiment 1: Multilayer Perceptron (MLP)*

The baseline performance of the Multilayer Perceptron (MLP) model was evaluated using the resampled credit card transaction data, which was balanced through the Borderline-SMOTE technique. The training and validation loss over epochs is depicted in Figure 10 in Appendix A. The training process demonstrated a consistent decline with stabilization occurring after a few epochs, indicating effective learning and generalization.

Table 4 presents a comparative analysis of the three models for fraud detection. The MLP model demonstrated strong baseline performance, achieving high precision, recall, and AUC-ROC score, indicating its ability to differentiate between fraudulent and legitimate transactions. The results of this experiment align closely with existing literature on the European Cardholders dataset, which employs an MLP classifier. These high-performance metrics are consistent with the findings reported in similar studies, underscoring the model's efficacy in this domain (Misra et al., 2020).

Table 4: Comparative Analysis of Three Models for Fraud Detection

Model	**Precision**	**Recall**	**F1-Score**	**ROC AUC Score**
MLP	0.77	0.82	0.79	0.980
Autoencoder + MLP	0.64	0.14	0.23	0.940
VAE + MLP	0.82	0.66	0.73	0.957

This experiment showed that the MLP model, when combined with Borderline SMOTE for class balancing, is highly effective for credit card fraud detection, achieving high precision and recall while maintaining excellent overall accuracy. As shown in Figure 4, the confusion matrix shows a low number of misclassifications, with only 24 false positives and 18 false negatives. Considering the focus of this thesis on minimizing false negatives (FN), MLP would be the preferred choice to improve fraud detection performance and prevent financial losses.

The precision-recall curve further validates the model's robustness, showing high precision even as recall increases. The ROC curve illustrates the true positive rate against the false positive rate for the MLP model,

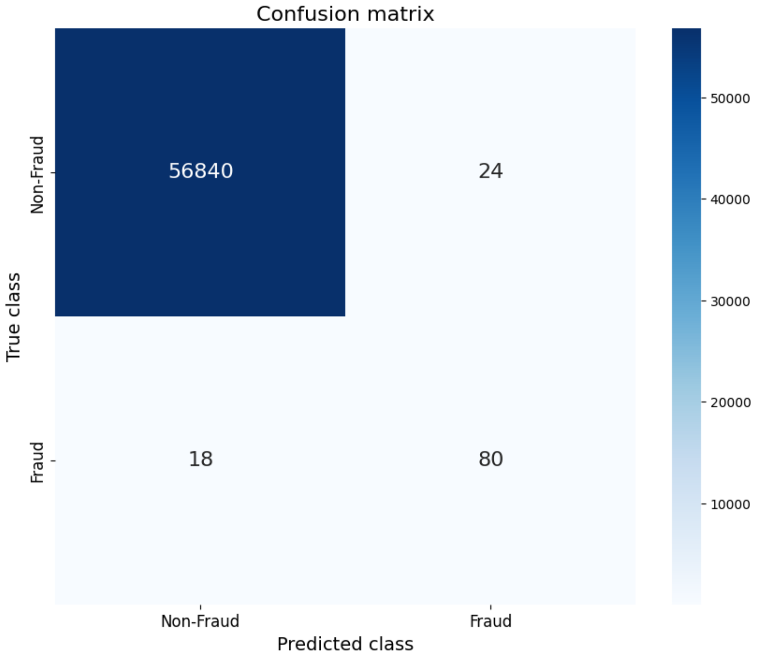

Figure 4: Confusion Matrix of the Best Performing Model

indicating excellent discriminative ability between fraudulent and non-fraudulent transactions (Figures 11, 12 in Appendix A).

The subsequent section examines whether implementing autoencoders enhances the classifier's performance in this task.

6.2 *Results of Experiment 2: Standard Autoencoder + Multilayer Perceptron (MLP)*

A standard autoencoder was integrated with the MLP classifier in the second experiment. The performance of this integrated model was compared against the baseline MLP model.

The autoencoder was trained on resampled data using Borderline-SMOTE. The encoded features obtained from the autoencoder were then used as training data for the MLP classifier. Table 4 demonstrates that the integration of the Standard Autoencoder with MLP resulted in a decrease in both recall and F1-score. This indicates that the latent representations generated by the autoencoder did not significantly improve the performance of MLP in this particular situation. The autoencoder integrated model's performance was inferior to the VAE integrated model and the standalone MLP model.

The confusion matrix of the model depicted in Figure 13, as presented in Appendix A, reveals that the model struggles with identifying fraudulent transactions, as indicated by the low recall and F1 score. Figure 14 illustrates the precision-recall curve, which offers further evidence of a significant decline in precision as recall increases. The model initially exhibits high precision; however, as the recall rate increases, the precision drops, suggesting that it faces difficulties in accurately detecting all instances of fraud.

Table 5: Performance Evaluation of Autoencoder Models Considering Temporal Transaction Differences

Model	Precision	Recall	F1-Score	AUC ROC Score
Autoencoder + MLP				
Short Time Difference	0.0203	0.8387	0.0396	0.9512
Long Time Difference	0.0327	0.8281	0.0629	0.9482
VAE + MLP				
Short Time Difference	0.0266	0.8925	0.0517	0.9499
Long Time Difference	0.0643	0.9062	0.1201	0.9390

Table 6: Performance Assessment of Autoencoder Models in Relation to Transaction Amounts

Model	Precision	Recall	F1-Score	AUC ROC Score
Autoencoder + MLP				
Low Amount Transactions	0.0174	0.8852	0.0341	0.9525
High Amount Transactions	0.0412	0.9459	0.0790	0.9682
VAE + MLP				
Low Amount Transactions	0.0270	0.9180	0.0524	0.9531
High Amount Transactions	0.0281	0.8378	0.0544	0.9271

Tables 5 and 6 display the experimental results for sub-questions using the segmented test data. According to Table 5, the performances of the two models varied based on temporal transaction differences. While the precision of the long-time difference is slightly better compared to the short-time difference, it still remains low. The high recall in both time differences suggests that the model is effective at identifying fraudulent transactions among the total fraudulent transactions. This situation demonstrates a trade-off where recall was prioritized over precision.

The results of the MLP classifier with integrated autoencoder for various transaction amounts are displayed in Table 6. The model achieved better performance scores on the high amount transactions in terms of better recall and improved precision. In general, this model shows slightly better AUC-ROC sores for high-amount transactions and short-time differences.

The findings from this experiment suggest that MLP combined with autoencoder is effective in certain aspects; however, there is still room for improvement in the detection of fraudulent transactions. The following section investigates whether the MLP classifier could be enhanced by the Variational Autoencoder (VAE).

6.3 *Results of Experiment 3: Variational Autoencoder (VAE) + Multilayer Perceptron (MLP)*

The MLP classifier was integrated with a Variational Autoencoder (VAE) in the third experiment. The performance metrics for the VAE + MLP model, as illustrated in Table 4, indicate a significant improvement over the standard autoencoder model and competitive performance compared to the standalone MLP model.

The VAE was trained on resampled data using the Borderline-SMOTE technique. The reconstruction error was utilized to detect anomalies, specifically fraudulent transactions. The reconstruction error refers to the difference between the original input and its reconstruction by the autoencoder. Figures 5 and 6 illustrate the reconstruction error for normal and fraudulent transactions using variational autoencoders. The histogram displayed in Figure 5 indicates that the majority of normal transactions have low reconstruction errors, which are primarily concentrated around zero. The small number of transactions with higher reconstruction errors are likely outliers, but they still fall within a low range. Figure 6 displays that the reconstruction errors associated with fraudulent transactions are noticeably higher and have a wider distribution. This indicates that the VAE encounters difficulties in accurately reconstructing fraudulent transactions, thus resulting in higher errors.

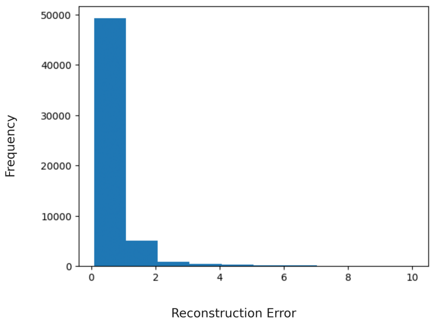

Figure 5: Reconstruction Error without Fraud: Variational Autoencoder

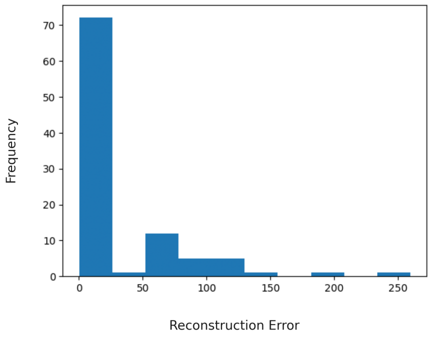

Figure 6: Reconstruction Error with Fraud: Variational Autoencoder

Furthermore, the reconstruction error plot in Figure 7 provides insights into the VAE's ability to reconstruct normal and fraudulent transactions, with a horizontal line indicating the threshold for anomaly detection. The threshold was set at 2.9, as shown by the red horizontal line in the plot. This threshold was chosen based on the distribution of reconstruction errors to distinguish between legitimate and fraudulent transactions effectively. The selected threshold approach was intended to represent the validity of the VAE in identifying fraudulent activities.

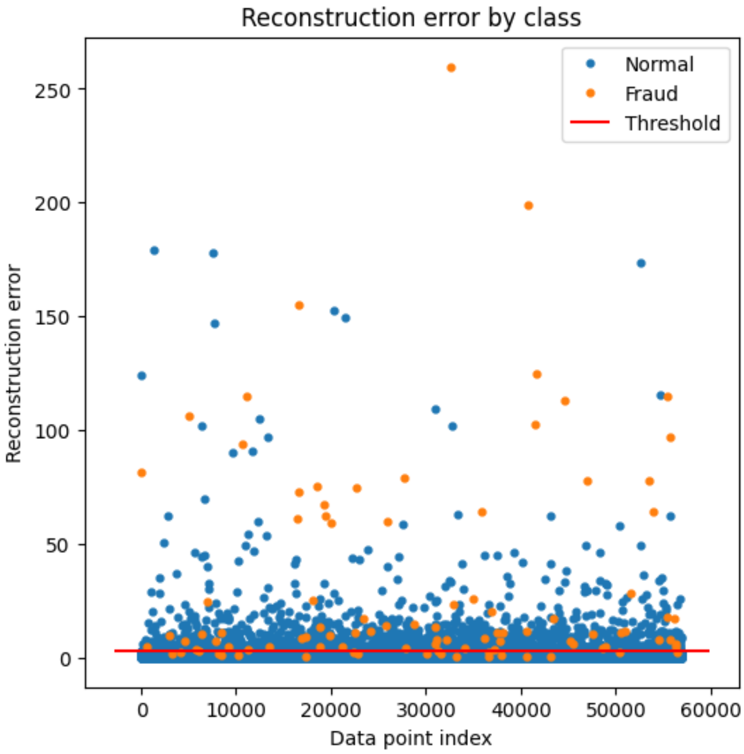

Figure 7: Reconstruction Error with Threshold: Variational Autoencoder

According to Table 4, the standalone MLP model exhibits a higher recall and slightly better F1-score and ROC-AUC score. Nevertheless, the VAE-integrated MLP model demonstrated higher precision. These findings suggest that while the MLP model had a higher overall success rate in identifying fraudulent transactions, the VAE integrated model had a higher level of precision in its fraud predictions, which is crucial for reducing the number of false positives.

Tables 5 and 6 reveal that both the amount of transactions and temporal differences have an impact on the performance of the model. Table 5 shows that the MLP combined with the VAE generally outperforms the model with the standard autoencoder across all metrics, especially for long-time differences, where it achieves the highest precision, recall, and F1 score. Table 6 illustrates that the VAE + MLP model consistently shows better recall and F1-Score for low-amount transactions, whereas the Autoencoder + MLP model demonstrates better performance for high-amount transactions in terms of AUC ROC Score.

The ROC curves for different amounts and time segments are shown in Figures 8 and 9. The model's consistent ability to maintain high AUC values across various segments demonstrates its reliability in practical fraud detection scenarios.

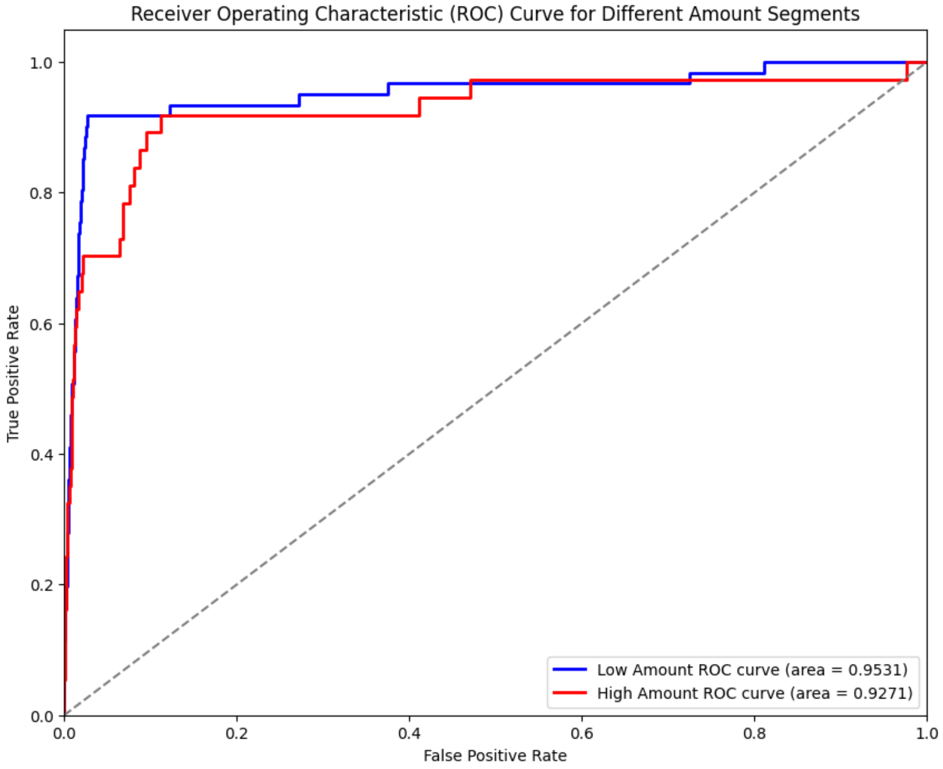

Figure 8: ROC Curve for Different Amount Segments

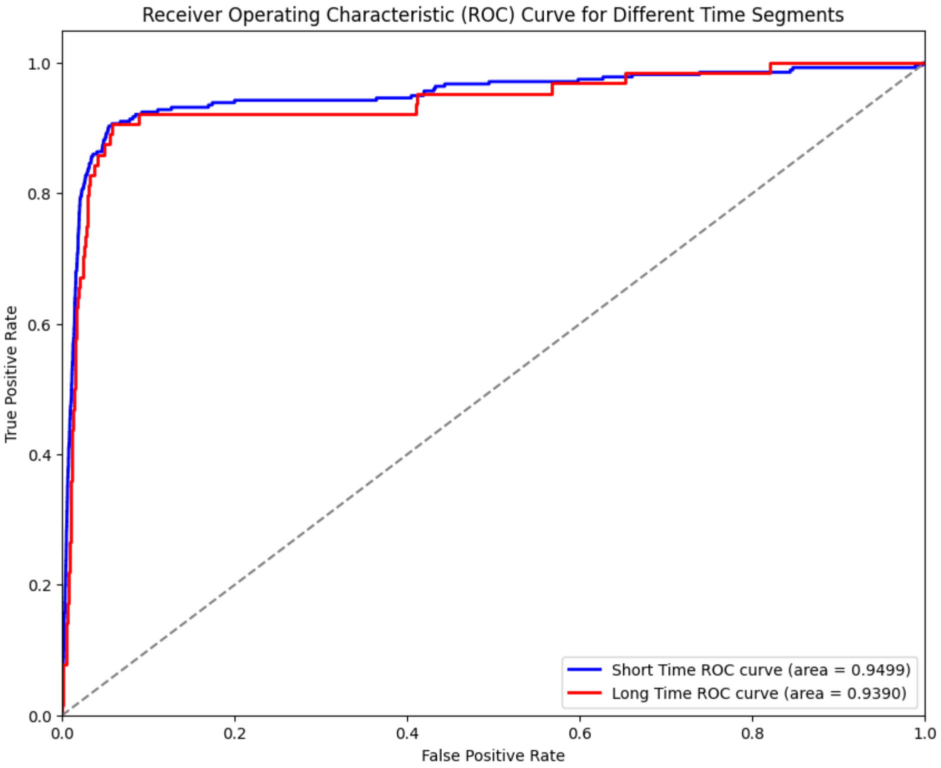

Figure 9: ROC Curve for Different Time Segments

The findings from Experiment 3 highlight the effectiveness of using a VAE over a standard autoencoder for preparing data for the MLP classifier in the domain of credit card fraud detection. According to Table 4, the enhanced precision of the model contributes to reducing false positives, thereby making VAE a valuable tool for minimizing financial losses due to fraud.

7 DISCUSSION

The primary goal of this thesis was to evaluate the effectiveness of different types of autoencoders in accurately identifying fraudulent transactions within the European Cardholders dataset. This research undertakes a comparative analysis of three models for credit card fraud detection: a standalone Multilayer Perceptron (MLP), an MLP combined with a standard autoencoder, and an MLP integrated with Variational Autoencoder (VAE).

Several challenges were encountered in this study, including the significant class imbalance between fraudulent and legitimate transactions. Additionally, the limited availability of diverse datasets for training and testing further constrained the generalizability of findings. Despite these challenges, the study provides valuable insights into the capabilities and limitations of using autoencoders for credit card fraud detection.

7.1 *Key Findings*

The results of this study indicated that the standalone MLP achieved high-performance metrics compared to the models using the autoencoders. The first experiment demonstrated that the MLP is highly effective in distinguishing between fraudulent and legitimate transactions. These results align closely with existing literature and they highlight the MLP's ability to identify fraudulent transactions while minimizing the false negative rates.

The training dataset exhibited an extreme class imbalance between fraudulent and non-fraudulent transactions. To tackle this issue, Borderline-SMOTE was employed to increase the number of fraudulent transactions in the training dataset. The use of the Borderline-SMOTE technique significantly contributed to enhancing the model's sensitivity to minority class. However, this technique could potentially affect the performance of the model due to certain drawbacks. Generating synthetic examples can occasionally undermine the quality of data and potentially lead to the introduction of noise or duplicates. Therefore, the class imbalance may still persist to some extent. This can have an impact on the precision and recall measures, causing them to be lower compared to the AUC-ROC score and accuracy.

Moreover, these results were significantly influenced by hyperparameter tuning and the complexity of the model. Although the current architecture demonstrated impressive performance, exploring more complex architectures could potentially achieve even better scores.

The second experiment aimed to determine if training the MLP classifier on encoded features, which are compressed representations of the training data, could effectively capture the most informative parts required for this classification task. The objective was to assess whether using the encoded output from a standard autoencoder for classification could potentially increase model performance by reducing dimensionality. Additionally, this reduction was also intended to help minimize noise from the Borderline-SMOTE technique and improve model performance in fraud detection.

Unfortunately, the integration of a standard autoencoder with the MLP classifier did not significantly enhance performance during the second experiment. The recall and F1 score exhibited a decrease when compared to the baseline model in the first experiment. This suggests that latent representations of the autoencoder were not particularly beneficial in this specific context. Indeed, the integration resulted in a decrease in the model's ability to accurately detect fraudulent transactions. Consequently, the features learned by the standard autoencoder were not sufficiently distinctive to enhance the detection capabilities of the MLP. The utilization of the same MLP model architecture might not have resulted in optimal integration with the standard autoencoder. Thus, it may be necessary to adjust the architecture of MLP in order to leverage the encoded features better.

This study proceeded under the assumption that both fraudulent and non-fraudulent transactions possess certain shared characteristics within themselves. As a result, the autoencoder was trained on both fraudulent and non-fraudulent transactions rather than exclusively on normal transactions. This approach may have confused the algorithm in its ability to recognize the distinctions between transactions, resulting in a decline in performance.

In addition, the Borderline-SMOTE may also introduce bias to the performance of the autoencoder. The generated synthetic examples may not accurately reflect the actual fraudulent transactions, which can affect the autoencoder's ability to learn meaningful features. The variables in this study were not analyzed based on their significance in this dataset; hence, utilizing an autoencoder to reconstruct the data did not provide a significant benefit.

In the third experiment, the VAE-integrated MLP model exhibited a notable improvement over the standard autoencoder model due to the VAE's probabilistic approach. As a result, the MLP was trained on more complex patterns in the data, leading to more accurate identification of fraudulent transactions. Consequently, it achieved better precision and recall and showed a competitive performance compared to the standalone MLP.

Compared to the second experiment, the incorporation of VAE reduced false negatives, which is crucial for preventing financial losses in fraud detection systems. However, the model's performance in predicting non-fraudulent transactions slightly decreased (Figures 13, 15, Appendix A).

The findings from the third experiment prove that the enhanced feature extraction capabilities of VAEs likely contributed to better discrimination between fraudulent and non-fraudulent transactions. Several reasons could be behind this improvement, such as better integration with MLP, improved feature extraction, and effective dimensionality reduction. The collaboration between VAE and MLP could have strengthened the learning process and improved classification performance.

Additionally, it is important to consider the potential impact of Borderline-SMOTE on the VAE. Poor data quality or duplicate synthetic samples can lead to improper learning and performance. Nevertheless, the VAE's probabilistic framework likely mitigated these problems more effectively than standard autoencoders by modeling accurately the underlying data distribution.

7.2 *Limitations*

This thesis had several limitations that should be acknowledged. Firstly, the substantial class imbalance between fraudulent and legitimate transactions presented a major challenge. Despite the utilization of the Borderline-SMOTE technique, creating synthetic instances close to the borderline of the classes can potentially introduce noise or repetitions, which could undermine the data quality and lead to low model performance.

Secondly, the absence of varied datasets for both training and testing further restricted the applicability of the results. The reliance on the European Cardholders dataset may limit the relevance of the findings to other datasets with different distributions and fraud patterns. Additional datasets with diverse attributes could offer more comprehensive insights into the model's effectiveness in different scenarios.

Thirdly, the study did not conduct a thorough analysis of the importance of individual features due to the nature of the dataset. The dataset used in this study included numerical input attributes transformed by Principal Component Analysis (PCA). While anonymization protected sensitive user data, contextual details were lost. Therefore, the inability to analyze features based on their importance may have impacted the effectiveness of autoencoders in learning meaningful representations from the most discriminative characteristics for fraud detection. Implementing feature importance analysis could enhance the understanding of which features are most indicative of fraudulent activity, thereby potentially improving the autoencoders' performance.

Lastly, the hyperparameter tuning and complexity of the models were influenced by the time constraints and computational resources. By exploring more complex architectures and advanced optimization techniques, it is possible to achieve better results.

By acknowledging these limitations, future studies can focus on addressing these difficulties to further improve the effectiveness of credit card fraud detection systems using autoencoders.

7.3 *Relevance*

As highlighted in this thesis, limited studies have focused on the application of autoencoders for credit card fraud detection. By comparing different autoencoders integrated with MLP models, the research demonstrates the potential of VAE-integrated models to capture meaningful data representations, which can improve the accuracy of fraud detection.

The growing need for effective fraud detection systems in the financial sector emphasizes the significance of this study. Efficient fraud detection not only reduces financial losses but also promotes trust in financial institutions and credit card usage. Additionally, the ability to train models effectively while ensuring data privacy and security is crucial in this field. These findings are consistent with prior research that highlights the challenges and benefits of using anonymized data for machine learning applications.

Furthermore, this study explores advanced deep learning techniques and their integration, providing valuable insights into model selection and optimization. By highlighting the benefits and limitations of these approaches, this study offers a crucial resource for developing more effective fraud detection systems.

7.4 *Future Work*

There are multiple options for conducting further research. Firstly, it would be advantageous to investigate the incorporation of additional deep learning architectures with autoencoders, such as Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs), to capture deeper patterns in the data. These models have the potential to improve the robustness of fraud detection systems.

Furthermore, it is possible to identify the most significant features by incorporating methods such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-Agnostic Explanations). These techniques can enhance the model performance by highlighting which features are most influential in detecting fraud.

Additionally, It is essential to validate the generalizability of the models by testing them on diverse, real-world datasets from various financial institutions. Collaborations with industry partners to access such data could offer valuable insights into the practical feasibility of the models. Introducing more features, such as geolocation data, device information, and user behavior patterns, could improve the learning process of the models. Geolocation data can highlight unusual transaction locations, device information can detect unfamiliar devices, and user behavior patterns can identify anomalies in spending habits.

Finally, it is crucial to consider the quality of synthetic data generated by Borderline-SMOTE. One possible option to improve model performance is to investigate the use of advanced techniques, such as Generative Adversarial Networks (GANs), to create more realistic synthetic examples. This could improve the training process and performance of the models by ensuring higher-quality synthetic samples.

8 CONCLUSION

This section concludes with the answers to the research questions presented in the thesis.

RQ: "To what extent can different types of autoencoders effectively distinguish fraudulent credit card transactions from legitimate ones?"

To answer this primary research question, this thesis investigated the comparative performances of a simple Multilayer Perceptron (MLP), an MLP combined with a standard autoencoder, and an MLP integrated with a Variational Autoencoder (VAE) for credit card fraud detection.

The standalone MLP model outperformed the other models by achieving high precision, recall, and AUC-ROC scores. The Borderline-SMOTE technique was essential in handling the class imbalance and improving the model sensitivity to the minority class. This suggests that training the MLP with balanced features enables it to retain all relevant information for classification. In contrast, the standard autoencoder faced difficulties in preserving critical features for effective fraud detection. This could be attributed to the autoencoder's latent representation not accurately reflecting the original data. However, the VAE-integrated MLP model showed significant improvement over the standard autoencoder by capturing complex data patterns effectively due to its probabilistic approach.

In conclusion, while standard autoencoders are less effective for fraud detection, VAEs offer a better approach by enhancing data representation, especially when combined with Borderline-SMOTE. Nonetheless, a well-tuned MLP trained with data balanced by Borderline SMOTE remains highly effective for this task. Future work could investigate the potential of hybrid models that leverage the strengths of both VAEs and MLPs, as well as advanced balancing techniques to further improve fraud detection performance.

Sub-RQ1: "How does the transaction amount influence the autoencoders' ability to detect fraudulent transactions?"

To comprehensively examine the influence of transaction amounts on the performance of autoencoders, the test data was segmented into low and high transaction amounts. After training the autoencoders and the MLP classifier on the entire training dataset, the MLP was evaluated on each segment. These steps were undertaken to identify any correlations between transaction amounts and model performance.

Both models exhibited high recall and AUC-ROC scores for both low and high-amount transactions. The VAE combined with the MLP model demonstrated slightly better precision for low transactions while maintaining comparable recall and AUC ROC scores. Conversely, the standard autoencoder integrated model performed better in high-amount transactions in terms of AUC ROC, although its precision remained low. This variation implies that different types of autoencoders may be more suitable for varying transaction amounts.

However, transaction amount alone does not provide a sufficient basis for distinguishing between fraudulent and non-fraudulent transactions. Although the models demonstrated high recall, they suffered from low precision, indicating that many non-fraudulent transactions are misclassified as fraudulent. This underscores the necessity for incorporating additional features to improve the overall performance of fraud detection models.

Sub-RQ2: "How does the time difference between transactions affect the performance of autoencoders in detecting fraud?"

To investigate this question, the test data was divided into short and long intervals based on the time differences between consecutive transactions. The models were evaluated on these segments to determine any correlations between transaction timing and model performance.

The analysis showed that both models achieved high recall across short and long intervals, so they can effectively detect most fraudulent activities regardless of the time difference. The implementation of the Borderline-SMOTE technique played a crucial role in achieving high recall rates. However, both models suffered from low precision, implying a high rate of false positives where many legitimate transactions were incorrectly classified as fraudulent (Figures 16, 17, 18, 19 Appendix A). The VAE integrated model exhibited slightly better precision in long-time differences compared to the standard autoencoder model. As a result, VAE may be more effective in capturing patterns in transactions that occur over longer periods.

To conclude, while the VAE-integrated MLP model generally outperformed the standard autoencoder in terms of overall metrics, temporal differences alone are insufficient for accurately distinguishing between fraudulent and non-fraudulent transactions. The findings underscore the necessity of incorporating additional features for effective fraud detection systems.

REFERENCES

Abd El Naby, A., Hemdan, E. E.-D., & El-Sayed, A. (2021). Deep learning approach for credit card fraud detection. *2021 International Conference on Electronic Engineering (ICEEM)*, 1–5.

Abd El-Naby, A., Hemdan, E. E.-D., & El-Sayed, A. (2023). An efficient fraud detection framework with credit card imbalanced data in financial services. *Multimedia Tools and Applications*, *82*(3), 4139–4160.

Aftabi, S. Z., Ahmadi, A., & Farzi, S. (2023). Fraud detection in financial statements using data mining and gan models. *Expert Systems with Applications*, *227*, 120144.

Aguiar, G., Krawczyk, B., & Cano, A. (2023). A survey on learning from imbalanced data streams: Taxonomy, challenges, empirical study, and reproducible experimental framework. *Machine learning*, 1–79.

Ahmad, H., Kasasbeh, B., Aldabaybah, B., & Rawashdeh, E. (2023). Class balancing framework for credit card fraud detection based on clustering and similarity-based selection (sbs). *International Journal of Information Technology*, *15*(1), 325–333.

Alam, B. R., Khatun, M. S., Taslim, M., Hossain, M. A., et al. (2022). Handling class imbalance in credit card fraud using various sampling techniques. *American Journal of Multidisciplinary Research and Innovation*, *1*(4), 160–168.

Alamri, M., & Ykhlef, M. (2022). Survey of credit card anomaly and fraud detection using sampling techniques. *Electronics*, *11*(23), 4003.

Alamri, M., & Ykhlef, M. (2024). Hybrid undersampling and oversampling for handling imbalanced credit card data. *IEEE Access*.

Alazizi, A., Habrard, A., Jacquenet, F., He-Guelton, L., & Oblé, F. (2020). Dual sequential variational autoencoders for fraud detection. *Advances in Intelligent Data Analysis XVIII: 18th International Symposium on Intelligent Data Analysis, IDA 2020, Konstanz, Germany, April 27–29, 2020, Proceedings 18*, 14–26.

Ali, A., Abd Razak, S., Othman, S. H., Eisa, T. A. E., Al-Dhaqm, A., Nasser, M., Elhassan, T., Elshafie, H., & Saif, A. (2022). Financial fraud detection based on machine learning: A systematic literature review. *Applied Sciences*, *12*(19), 9637.

Ali, H., Salleh, M. M., Saedudin, R., Hussain, K., & Mushtaq, M. F. (2019). Imbalance class problems in data mining: A review. *Indonesian Journal of Electrical Engineering and Computer Science*, *14*(3), 1560–1571.

Almhaithawi, D., Jafar, A., & Aljnidi, M. (2020). Example-dependent cost-sensitive credit cards fraud detection using smote and bayes minimum risk. *SN Applied Sciences*, *2*, 1–12.

Al-Shabi, M. (2019). Credit card fraud detection using autoencoder model in unbalanced datasets. *Journal of Advances in Mathematics and Computer Science*, *33*(5), 1–16.

An, J., & Cho, S. (2015). Variational autoencoder based anomaly detection using reconstruction probability. *Special lecture on IE*, *2*(1), 1–18.

Anh, N. T. N., Khanh, T. Q., Dat, N. Q., Amouroux, E., & Solanki, V. K. (2020). Fraud detection via deep neural variational autoencoder oblique random forest. *2020 IEEE-HYDCON*, 1–6.

Asha, R., & KR, S. K. (2021). Credit card fraud detection using artificial neural network. *Global Transitions Proceedings*, *2*(1), 35–41.

Awosika, T., Shukla, R. M., & Pranggono, B. (2024). Transparency and privacy: The role of explainable ai and federated learning in financial fraud detection. *IEEE Access*.

Baesens, B., Höppner, S., & Verdonck, T. (2021). Data engineering for fraud detection. *Decision Support Systems*, *150*, 113492.

Benchaji, I., Douzi, S., El Ouahidi, B., & Jaafari, J. (2021). Enhanced credit card fraud detection based on attention mechanism and lstm deep model. *Journal of Big Data*, *8*, 1–21.

Bengio, Y., Courville, A. C., & Vincent, P. (2012). Unsupervised feature learning and deep learning: A review and new perspectives. *CoRR, abs/1206.5538*, *1*(2665), 2012.

Bin Sulaiman, R., Schetinin, V., & Sant, P. (2022). Review of machine learning approach on credit card fraud detection. *Human-Centric Intelligent Systems*, *2*(1-2), 55–68.

Carcillo, F., Le Borgne, Y.-A., Caelen, O., & Bontempi, G. (2018). Streaming active learning strategies for real-life credit card fraud detection: Assessment and visualization. *International Journal of Data Science and Analytics*, *5*, 285–300.

Caroline Cynthia, P., & Thomas George, S. (2021). An outlier detection approach on credit card fraud detection using machine learning: A comparative analysis on supervised and unsupervised learning. *Intelligence in Big Data Technologies—Beyond the Hype: Proceedings of ICBDCC 2019*, 125–135.

Chalapathy, R., Menon, A. K., & Chawla, S. (2018). Anomaly detection using one-class neural networks. *arXiv preprint arXiv:1802.06360*.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, *41*(3), 1–58.

Chaquet-Ulldemolins, J., Gimeno-Blanes, F.-J., Moral-Rubio, S., Muñoz-Romero, S., & Rojo-Álvarez, J.-L. (2022). On the black-box challenge for fraud detection using machine learning (ii): Nonlinear analysis through interpretable autoencoders. *Applied Sciences*, *12*(8), 3856.

Chen, C. P., & Zhang, C.-Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on big data. *Information sciences*, *275*, 314–347.

Chen, Z., Yeo, C. K., Lee, B. S., & Lau, C. T. (2018). Autoencoder-based network anomaly detection. *2018 Wireless telecommunications symposium (WTS)*, 1–5.

Chen, Z., Soliman, W. M., Nazir, A., & Shorfuzzaman, M. (2021). Variational autoencoders and wasserstein generative adversarial networks for improving the anti-money laundering process. *IEEE Access*, *9*, 83762–83785.

Cheng, D., Xiang, S., Shang, C., Zhang, Y., Yang, F., & Zhang, L. (2020). Spatio-temporal attention-based neural network for credit card fraud detection. *Proceedings of the AAAI conference on artificial intelligence*, *34*(01), 362–369.

Chollet, F., et al. (2015). *Keras*. https://github.com/fchollet/keras

Chow, J. K., Su, Z., Wu, J., Tan, P. S., Mao, X., & Wang, Y.-H. (2020). Anomaly detection of defects on concrete structures with the convolutional autoencoder. *Advanced Engineering Informatics*, *45*, 101105.

Dal Pozzolo, A., Caelen, O., Johnson, R. A., & Bontempi, G. (2015). Calibrating probability with undersampling for unbalanced classification. *2015 IEEE symposium series on computational intelligence*, 159–166.

Davidson, R. H. (2022). Who did it matters: Executive equity compensation and financial reporting fraud. *Journal of Accounting and Economics*, *73*(2-3), 101453.

De La Bourdonnaye, F., & Daniel, F. (2022). Evaluating resampling methods on a real-life highly imbalanced online credit card payments dataset. *arXiv preprint arXiv:2206.13152*.

Ding, Y., Kang, W., Feng, J., Peng, B., & Yang, A. (2023). Credit card fraud detection based on improved variational autoencoder generative adversarial network. *IEEE Access*.

Du, H., Lv, L., Guo, A., & Wang, H. (2023). Autoencoder and lightgbm for credit card fraud detection problems. *Symmetry*, *15*(4), 870.

Dubey, S. C., Mundhe, K. S., & Kadam, A. A. (2020). Credit card fraud detection using artificial neural network and backpropagation. *2020 4th international conference on intelligent computing and control systems (ICICCS)*, 268–273.

Ebenuwa, S. H., Sharif, M. S., Alazab, M., & Al-Nemrat, A. (2019). Variance ranking attributes selection techniques for binary classification problem in imbalance data. *IEEE access*, *7*, 24649–24666.

Fanai, H., & Abbasimehr, H. (2023). A novel combined approach based on deep autoencoder and deep classifiers for credit card fraud detection. *Expert Systems with Applications*, *217*, 119562.

Fernández, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). Smote for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research*, *61*, 863–905.

Fiore, U., De Santis, A., Perla, F., Zanetti, P., & Palmieri, F. (2019). Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. *Information Sciences*, *479*, 448–455.

Fournier, Q., & Aloise, D. (2019). Empirical comparison between autoencoders and traditional dimensionality reduction methods. *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, 211–214.

Gangwar, A. K., & Ravi, V. (2019). Wip: Generative adversarial network for oversampling data in credit card fraud detection. *Information Systems Security: 15th International Conference, ICISS 2019, Hyderabad, India, December 16–20, 2019, Proceedings 15*, 123–134.

Georgieva, S., Markova, M., & Pavlov, V. (2019). Using neural network for credit card fraud detection. *AIP Conference Proceedings*, *2159*(1).

Ghaleb, F. A., Saeed, F., Al-Sarem, M., Qasem, S. N., & Al-Hadhrami, T. (2023). Ensemble synthesized minority oversampling based generative adversarial networks and random forest algorithm for credit card fraud detection. *IEEE Access*.

Habeeb, R. A. A., Nasaruddin, F., Gani, A., Hashem, I. A. T., Ahmed, E., & Imran, M. (2019). Real-time big data processing for anomaly detection: A survey. *International Journal of Information Management*, *45*, 289–307.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert systems with applications*, *73*, 220–239.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., . . . Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hejazi, M., & Singh, Y. P. (2013). One-class support vector machines approach to anomaly detection. *Applied Artificial Intelligence*, *27*(5), 351–366.

Hilal, W., Gadsden, S. A., & Yawney, J. (2022). Financial fraud: A review of anomaly detection techniques and recent advances. *Expert systems With applications*, *193*, 116429.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *science*, *313*(5786), 504–507.

Hung, S.-K., & Gan, J. Q. (2021). Augmentation of small training data using gans for enhancing the performance of image classification. *2020 25th international conference on pattern recognition (ICPR)*, 3350–3356.

Ibrahim, B. I., Nicolae, D. C., Khan, A., Ali, S. I., & Khattak, A. (2020). Vae-gan based zero-shot outlier detection. *Proceedings of the 2020 4th international symposium on computer science and intelligent control*, 1–5.

Ileberi, E., Sun, Y., & Wang, Z. (2021). Performance evaluation of machine learning methods for credit card fraud detection using smote and adaboost. *IEEE Access*, *9*, 165286–165294.

Injadat, M., Salo, F., Nassif, A. B., Essex, A., & Shami, A. (2018). Bayesian optimization with machine learning algorithms towards anomaly detection. *2018 IEEE global communications conference (GLOBECOM)*, 1–6.

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, *6*(1), 1–54.

Kasasbeh, B., Aldabaybah, B., & Ahmad, H. (2022). Multilayer perceptron artificial neural networks-based model for credit card fraud detection. *Indonesian Journal of Electrical Engineering and Computer Science*, *26*(1), 362–373.

Kazemi, Z., & Zarrabi, H. (2017). Using deep networks for fraud detection in the credit card transactions. *2017 IEEE 4th International conference on knowledge-based engineering and innovation (KBEI)*, 0630–0633.

Kim, E., Lee, J., Shin, H., Yang, H., Cho, S., Nam, S.-k., Song, Y., Yoon, J.-a., & Kim, J.-i. (2019). Champion-challenger analysis for credit card fraud detection: Hybrid ensemble and deep learning. *Expert Systems with Applications*, *128*, 214–224.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., & Willing, C. Jupyter notebooks – a publishing format for reproducible computational workflows (F. Loizides & B. Schmidt, Eds.). In: *Positioning and power in academic publishing: Players, agents and agendas* (F. Loizides & B. Schmidt, Eds.). Ed. by Loizides, F., & Schmidt, B. IOS Press. 2016, 87–90.

Kumar, S. N. P. (2022). *Improving fraud detection in credit card transactions using autoencoders and deep neural networks* [Doctoral dissertation, The George Washington University].

Le, T. (2022). A comprehensive survey of imbalanced learning methods for bankruptcy prediction. *IET Communications*, *16*(5), 433–441.

Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, *18*(17), 1–5. http://jmlr.org/papers/v18/16-365.html

Lin, T.-H., & Jiang, J.-R. (2020). Anomaly detection with autoencoder and random forest. *2020 International Computer Symposium (ICS)*, 96–99.

Lin, T.-H., & Jiang, J.-R. (2021). Credit card fraud detection with autoencoder and probabilistic random forest. *Mathematics*, *9*(21), 2683.

Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., & Alsaadi, F. E. (2017). A survey of deep neural network architectures and their applications. *Neurocomputing*, *234*, 11–26.

López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information sciences*, *250*, 113–141.

Lucas, Y., & Jurgovsky, J. (2020). Credit card fraud detection using machine learning: A survey. *arXiv preprint arXiv:2010.06479*.

Lunghi, D., Paldino, G. M., Caelen, O., & Bontempi, G. (2023). An adversary model of fraudsters' behavior to improve oversampling in credit card fraud detection. *IEEE access*, *11*, 136666–136679.

Majhi, S. K., Bhatachharya, S., Pradhan, R., & Biswal, S. (2019). Fuzzy clustering using salp swarm algorithm for automobile insurance fraud detection. *Journal of Intelligent & Fuzzy Systems*, *36*(3), 2333–2344.

Makki, S. (2019, December). *An Efficient Classification Model for Analyzing Skewed Data to Detect Frauds in the Financial Sector* (Publication No. 2019LYSE1339) [Theses]. Université de Lyon ; Université Libanaise. https://theses.hal.science/tel-02457134

Makki, S., Assaghir, Z., Taher, Y., Haque, R., Hacid, M.-S., & Zeineddine, H. (2019). An experimental study with imbalanced classification approaches for credit card fraud detection. *IEEE Access*, *7*, 93010–93022.

Malini, N., & Pushpa, M. (2017). Analysis on credit card fraud identification techniques based on knn and outlier detection. *2017 third international conference on advances in electrical, electronics, information, communication and bio-informatics (AEEICB)*, 255–258.

Mehrotra, K. G., Mohan, C. K., Huang, H., Mehrotra, K. G., Mohan, C. K., & Huang, H. (2017). *Anomaly detection*. Springer.

Mishra, M. K., & Dash, R. (2014). A comparative study of chebyshev functional link artificial neural network, multi-layer perceptron and decision tree for credit card fraud detection. *2014 International Conference on Information Technology*, 228–233.

Misra, S., Thakur, S., Ghosh, M., & Saha, S. K. (2020). An autoencoder based model for detecting fraudulent credit card transaction. *Procedia Computer Science*, *167*, 254–262.

Moumeni, L., Saber, M., Slimani, I., Elfarissi, I., & Bougroun, Z. (2022). Machine learning for credit card fraud detection. *WITS 2020: Proceedings of the 6th International Conference on Wireless Technologies, Embedded, and Intelligent Systems*, 211–221.

Mrozek, P., Panneerselvam, J., & Bagdasar, O. (2020). Efficient resampling for fraud detection during anonymised credit card transactions with unbalanced datasets. *2020 IEEE/ACM 13th International Conference on Utility and Cloud Computing (UCC)*, 426–433.

Ngai, E. W., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision support systems*, *50*(3), 559–569.

Niu, X., Wang, L., & Yang, X. (2019). A comparison study of credit card fraud detection: Supervised versus unsupervised. *arXiv preprint arXiv:1904.10604*.

Obimbo, C., Mand, D., & Singh, S. (2021). Oversampling techniques in machine learning detection of credit card fraud. *Journal of Internet Technology and Secured Transactions (JITST)*, *9*, 741–746.

Ogwueleka, F. N. (2011). Data mining application in credit card fraud detection system. *Journal of Engineering Science and Technology*, *6*(3), 311–322.

Ouedraogo, A.-F., Heuchenne, C., Nguyen, Q.-T., & Tran, H. (2021). Data-driven approach for credit card fraud detection with autoencoder and one-class classification techniques. *Advances in Production Management Systems. Artificial Intelligence for Sustainable and Resilient Production Systems: IFIP WG 5.7 International Conference, APMS 2021, Nantes, France, September 5–9, 2021, Proceedings, Part I*, 31–38.

Ounacer, S., El Bour, H. A., Oubrahim, Y., Ghoumari, M. Y., & Azzouazi, M. (2018). Using isolation forest in anomaly detection: The case of credit card transactions. *Periodicals of Engineering and Natural Sciences*, *6*(2), 394–400.

Pandey, A., Bhatt, D., & Bhowmik, T. (2020). Limitations and applicability of gans in banking domain. *ADGN@ ECAI*.

Pang, G., Shen, C., Cao, L., & Hengel, A. V. D. (2021). Deep learning for anomaly detection: A review. *ACM computing surveys (CSUR)*, *54*(2), 1–38.

Patidar, R., Sharma, L., et al. (2011). Credit card fraud detection using neural network. *International Journal of Soft Computing and Engineering (IJSCE)*, *1*(32-38).

Patil, S., Nemade, V., & Soni, P. K. (2018). Predictive modelling for credit card fraud detection using data analytics. *Procedia computer science*, *132*, 385–395.

Patra, G. R., Iyer, S. P., Satyaprakash, S., & Mohanty, M. N. (2023). A class balancing based machine learning approach for fraudulent credit card transaction detection. *2023 1st International Conference on Circuits, Power and Intelligent Systems (CCPIS)*, 1–6.

Paula, E. L., Ladeira, M., Carvalho, R. N., & Marzagão, T. (2016). Deep learning anomaly detection as support fraud investigation in brazilian exports and anti-money laundering. *2016 15th ieee international conference on machine learning and applications (icmla)*, 954–960.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pillai, T. R., Hashem, I. A. T., Brohi, S. N., Kaur, S., & Marjani, M. (2018). Credit card fraud detection using deep learning technique. *2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA)*, 1–6.

Porwal, U., & Mukund, S. (2018). Credit card fraud detection in e-commerce: An outlier detection approach. *arXiv preprint arXiv:1811.02196*.

Prasetiyo, B., Muslim, M., Baroroh, N., et al. (2021). Evaluation performance recall and f2 score of credit card fraud detection unbalanced dataset using smote oversampling technique. *Journal of physics: conference series*, *1918*(4), 042002.

Pumsirirat, A., & Liu, Y. (2018). Credit card fraud detection using deep learning based on auto-encoder and restricted boltzmann machine. *International Journal of advanced computer science and applications*, *9*(1).

Raza, M., & Qayyum, U. (2019). Classical and deep learning classifiers for anomaly detection. *2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, 614–618.

Razooqi, T., Khurana, P., Raahemifar, K., & Abhari, A. (2016). Credit card fraud detection using fuzzy logic and neural network. *Proceedings of the 19th Communications & Networking Symposium*, 1–5.

Renström, M., & Holmsten, T. (2018). Fraud detection on unlabeled data with unsupervised machine learning.

Rezapour, M. (2019). Anomaly detection using unsupervised methods: Credit card fraud case study. *International Journal of Advanced Computer Science and Applications*, *10*(11).

Riffi, J., Mahraz, M. A., El Yahyaouy, A., Tairi, H., et al. (2020). Credit card fraud detection based on multilayer perceptron and extreme learning machine architectures. *2020 International Conference on Intelligent Systems and Computer Vision (ISCV)*, 1–5.

Rout, N., Mishra, D., & Mallick, M. K. (2018). Handling imbalanced data: A survey. *International Proceedings on Advances in Soft Computing, Intelligent Systems and Applications: ASISA 2016*, 431–443.

Roy, A., Sun, J., Mahoney, R., Alonzi, L., Adams, S., & Beling, P. (2018). Deep learning detecting fraud in credit card transactions. *2018 systems and information engineering design symposium (SIEDS)*, 129–134.

Saia, R., & Carta, S. (2019). Evaluating the benefits of using proactive transformed-domain-based techniques in fraud detection tasks. *Future Generation Computer Systems*, *93*, 18–32.

Sakurada, M., & Yairi, T. (2014). Anomaly detection using autoencoders with nonlinear dimensionality reduction. *Proceedings of the MLSDA 2014 2nd workshop on machine learning for sensory data analysis*, 4–11.

Schreyer, M., Sattarov, T., Borth, D., Dengel, A., & Reimer, B. (2017). Detection of anomalies in large scale accounting data using deep autoencoder networks. *arXiv preprint arXiv:1709.05254*.

Sehrawat, D., & Singh, Y. (2023). Auto-encoder and lstm-based credit card fraud detection. *SN Computer Science*, *4*(5), 557.

Shamsudin, H., Yusof, U. K., Jayalakshmi, A., & Khalid, M. N. A. (2020). Combining oversampling and undersampling techniques for imbalanced classification: A comparative study using credit card fraudulent transaction dataset. *2020 IEEE 16th international conference on control & automation (ICCA)*, 803–808.

Sharma, M. A., Raj, B. G., Ramamurthy, B., & Bhaskar, R. H. (2022). Credit card fraud detection using deep learning based on auto-encoder. *ITM Web of Conferences*, *50*, 01001.

Shen, J. (2021). Credit card fraud detection using autoencoder-based deep neural networks. *2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)*, 673–677.

Shirodkar, N., Mandrekar, P., Mandrekar, R. S., Sakhalkar, R., Kumar, K. C., & Aswale, S. (2020). Credit card fraud detection techniques–a survey. *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, 1–7.

Singh, A., Ranjan, R. K., & Tiwari, A. (2022). Credit card fraud detection under extreme imbalanced data: A comparative study of data-level algorithms. *Journal of Experimental & Theoretical Artificial Intelligence*, *34*(4), 571–598.

Singla, J., et al. (2020). A survey of deep learning based online transactions fraud detection systems. *2020 International Conference on Intelligent Engineering and Management (ICIEM)*, 130–136.

Sisodia, D. S., Reddy, N. K., & Bhandari, S. (2017). Performance evaluation of class balancing techniques for credit card fraud detection. *2017 IEEE International Conference on power, control, signals and instrumentation engineering (ICPCSI)*, 2747–2752.

Strelcenia, E., & Prakoonwit, S. (2023a). Improving classification performance in credit card fraud detection by using new data augmentation. *AI*, *4*(1), 172–198.

Strelcenia, E., & Prakoonwit, S. (2023b). A survey on gan techniques for data augmentation to address the imbalanced data issues in credit card fraud detection. *Machine Learning and Knowledge Extraction*, *5*(1), 304–329.

Sun, Y., Que, H., Cai, Q., Zhao, J., Li, J., Kong, Z., & Wang, S. (2022). Borderline smote algorithm and feature selection-based network anomalies detection strategy. *Energies*, *15*(13), 4751.

Sweers, T., Heskes, T., & Krijthe, J. (2018). Autoencoding credit card fraud. *Bachelor Thesis*.

Taha, A. A., & Malebary, S. J. (2020). An intelligent approach to credit card fraud detection using an optimized light gradient boosting machine. *IEEE Access*, *8*, 25579–25587.

Tingfei, H., Guangquan, C., & Kuihua, H. (2020). Using variational auto encoding in credit card fraud detection. *IEEE Access*, *8*, 149841–149853.

Tiwari, P., Mehta, S., Sakhuja, N., Kumar, J., & Singh, A. K. (2021). Credit card fraud detection using machine learning: A study. *arXiv preprint arXiv:2108.10005*.

Tyagi, S., & Mittal, S. (2020). Sampling approaches for imbalanced data classification problem in machine learning. *Proceedings of ICRIC 2019: Recent innovations in computing*, 209–221.

Vandewiele, G., Dehaene, I., Kovács, G., Sterckx, L., Janssens, O., Ongenae, F., De Backere, F., De Turck, F., Roelens, K., Decruyenaere, J., et al. (2021). Overly optimistic prediction results on imbalanced data: A case study of flaws and benefits when applying over-sampling. *Artificial Intelligence in Medicine*, *111*, 101987.

Varmedja, D., Karanovic, M., Sladojevic, S., Arsenovic, M., & Anderla, A. (2019). Credit card fraud detection-machine learning methods. *2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH)*, 1–5.

Wang, Y., Yao, H., Zhao, S., & Zheng, Y. (2015). Dimensionality reduction strategy based on auto-encoder. *Proceedings of the 7th International Conference on Internet Multimedia Computing and Service*, 1–4.

Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Xie, Y., Li, A., Hu, B., Gao, L., & Tu, H. (2023). A credit card fraud detection model based on multi-feature fusion and generative adversarial network. *Computers, Materials & Continua*, *76*(3).

Xuan, S., Liu, G., Li, Z., Zheng, L., Wang, S., & Jiang, C. (2018). Random forest for credit card fraud detection. *2018 IEEE 15th international conference on networking, sensing and control (ICNSC)*, 1–6.

Zamini, M., & Montazer, G. (2018). Credit card fraud detection using autoencoder based clustering. *2018 9th International Symposium on Telecommunications (IST)*, 486–491.

Zareapoor, M., Seeja, K., & Alam, M. A. (2012). Analysis on credit card fraud detection techniques: Based on certain design criteria. *International journal of computer applications*, *52*(3).

Zheng, P., Yuan, S., Wu, X., Li, J., & Lu, A. (2019). One-class adversarial nets for fraud detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 1286–1293.

Zhou, C., & Paffenroth, R. C. (2017). Anomaly detection with robust deep autoencoders. *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 665–674.

Zhu, H., Liu, G., Zhou, M., Xie, Y., Abusorrah, A., & Kang, Q. (2020). Optimizing weighted extreme learning machines for imbalanced classification and application to credit card fraud detection. *Neurocomputing*, *407*, 50–62.

Zhu, X., Ao, X., Qin, Z., Chang, Y., Liu, Y., He, Q., & Li, J. (2021). Intelligent financial fraud detection practices in post-pandemic era. *The Innovation*, *2*(4).

Zioviris, G., Kolomvatsos, K., & Stamoulis, G. (2022). Credit card fraud detection using a deep learning multistage model. *The Journal of Supercomputing*, *78*(12), 14571–14596.

Zou, H. (2021). Analysis of best sampling strategy in credit card fraud detection using machine learning. *Proceedings of the 2021 6th International Conference on Intelligent Information Technology*, 40–44.

Zou, J., Zhang, J., & Jiang, P. (2019). Credit card fraud detection using autoencoder neural network. *arXiv preprint arXiv:1908.11553*.

APPENDIX A

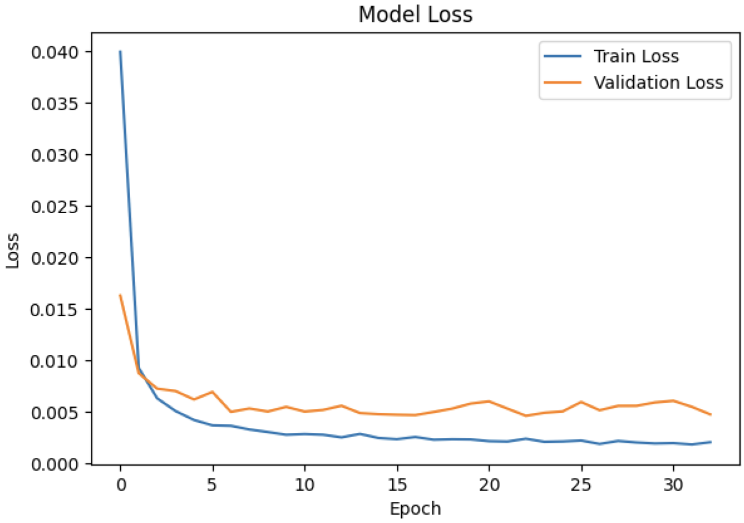

Figure 10: Model Loss: Multilayer Perceptron (MLP) with Borderline SMOTE

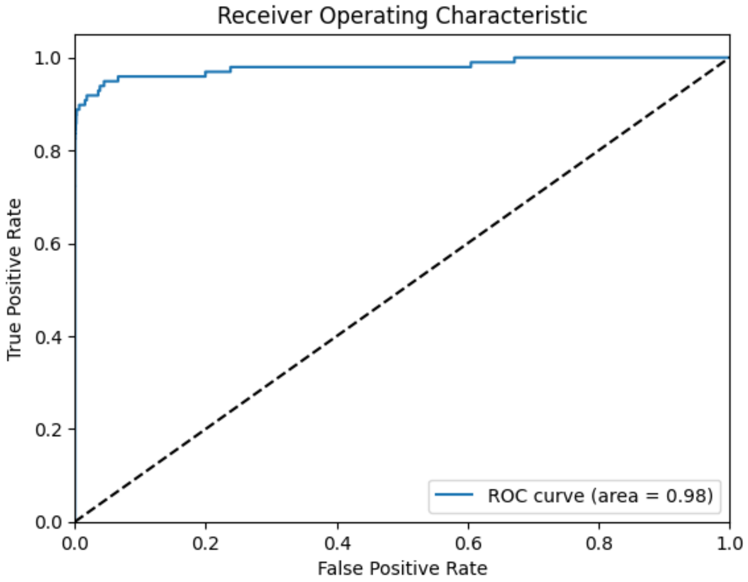

Figure 11: ROC Curve: Multilayer Perceptron (MLP) with Borderline SMOTE

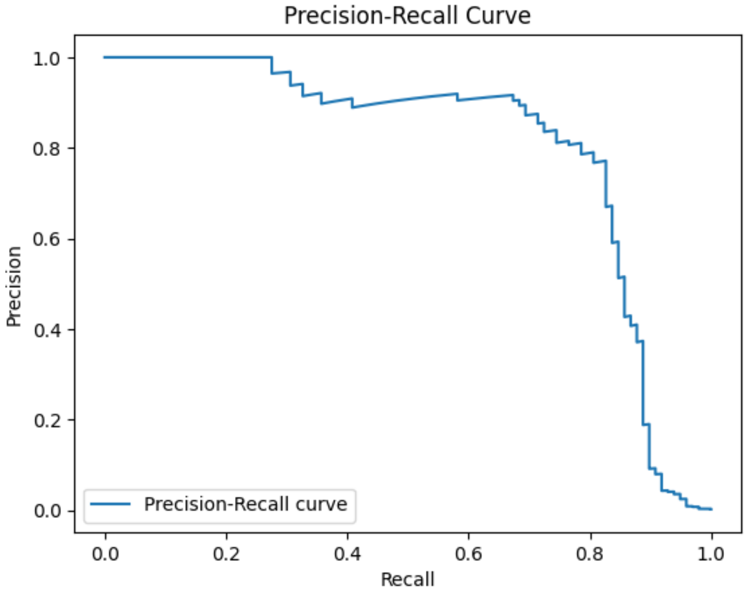

Figure 12: Precision-Recall Curve: Multilayer Perceptron (MLP) with Borderline SMOTE

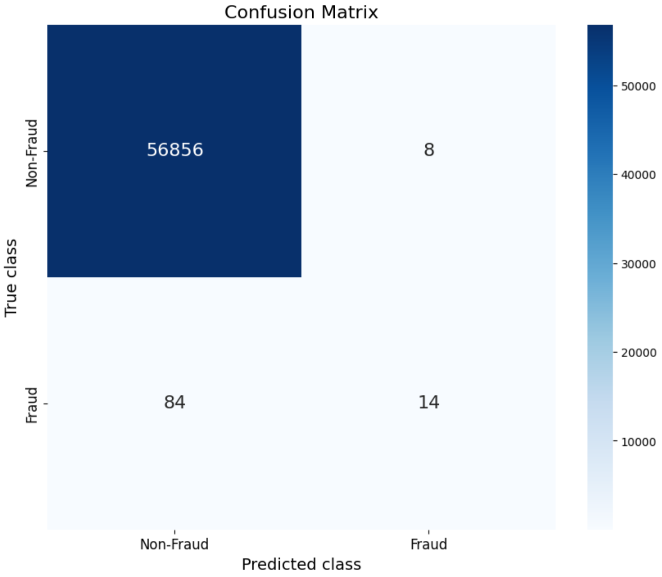

Figure 13: Confusion Matrix of Multilayer Perceptron with Integrated Autoencoder

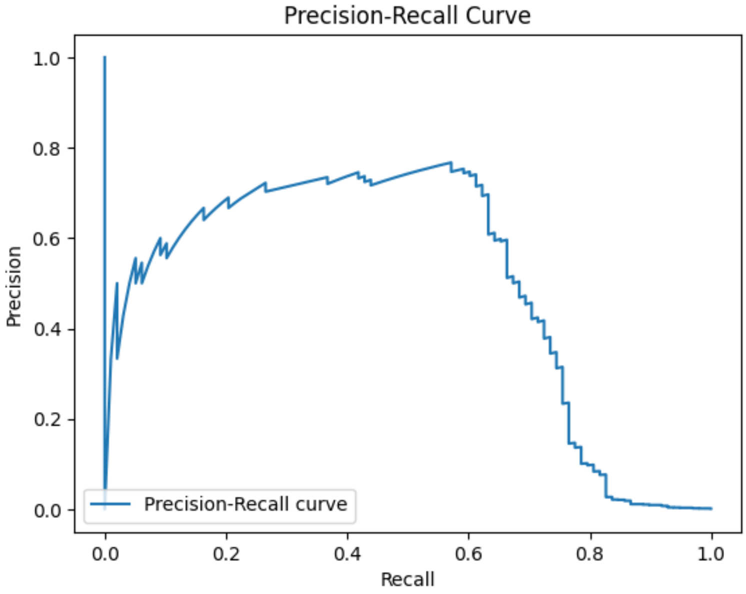

Figure 14: Precision-Recall Curve: Multilayer Perceptron (MLP) with Integrated Autoencoder

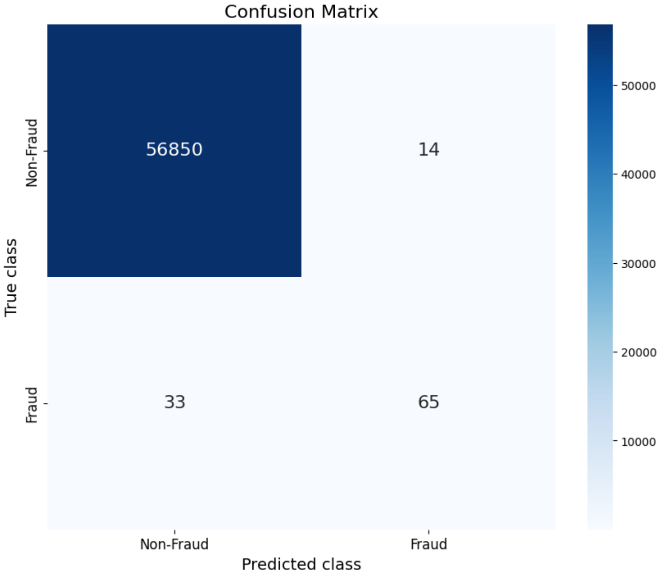

Figure 15: Confusion Matrix of Multilayer Perceptron with Integrated VAE

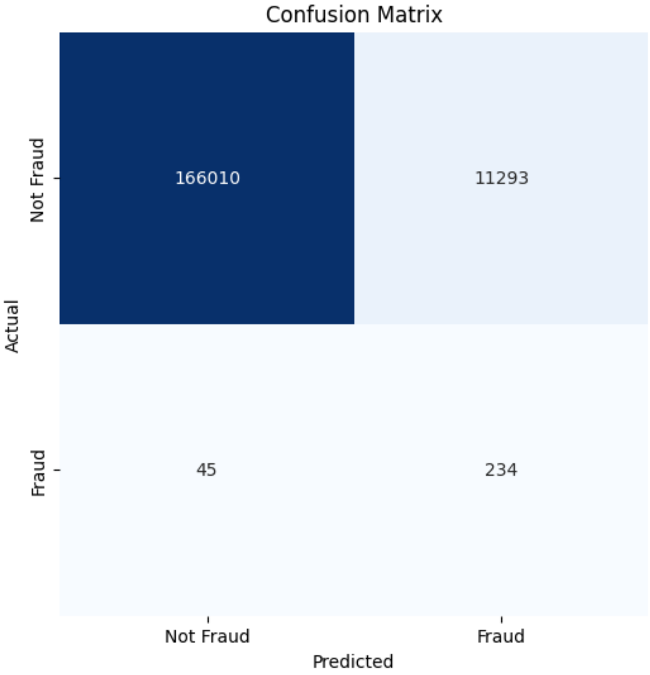

Figure 16: Confusion Matrix of MLP with Autoencoder on Short Time Difference Transactions

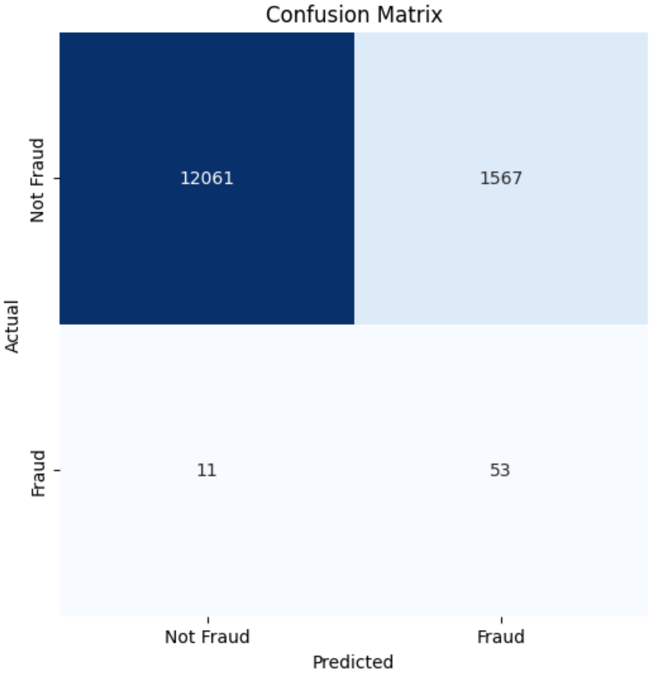

Figure 17: Confusion Matrix of MLP with Autoencoder on Long Time Difference Transactions

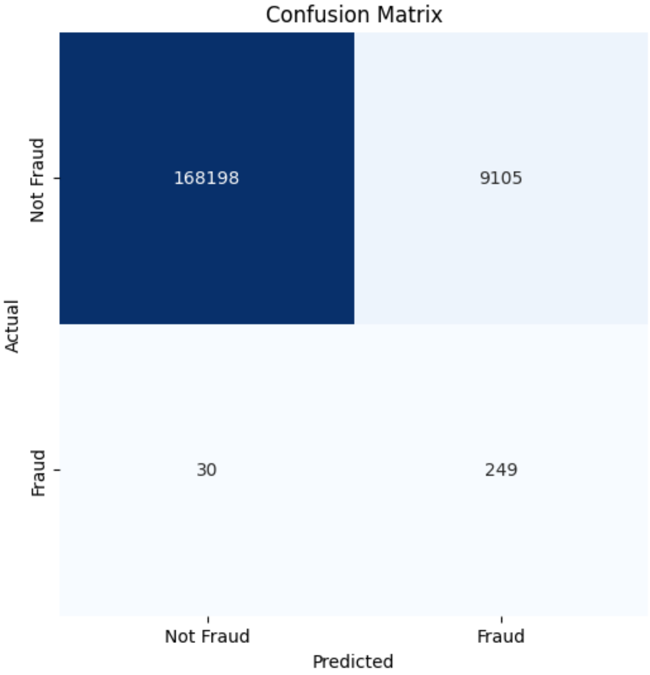

Figure 18: Confusion Matrix of MLP with VAE on Short Time Difference Transactions

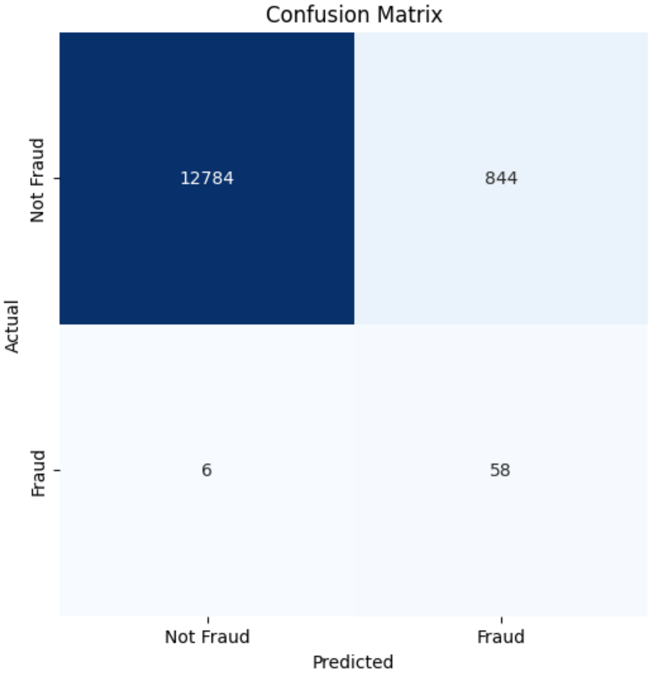

Figure 19: Confusion Matrix of MLP with VAE on Long Time Difference Transactions

APPENDIX B

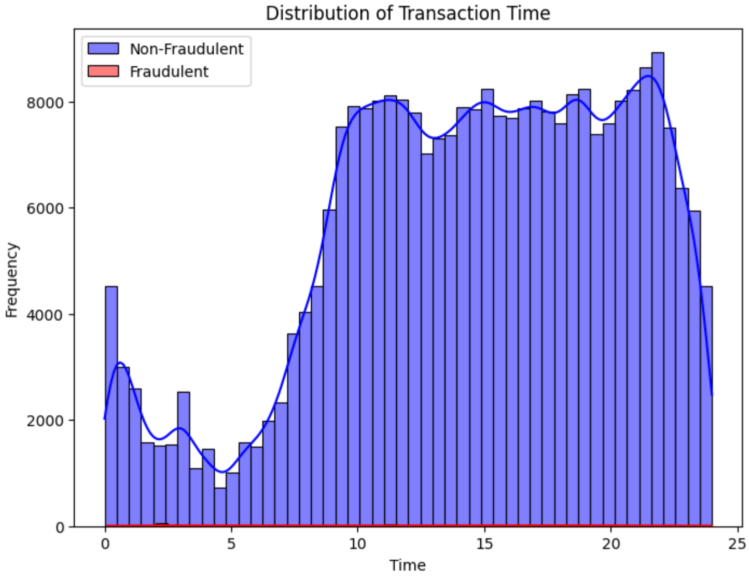

Figure 20: Distribution of Transaction Time

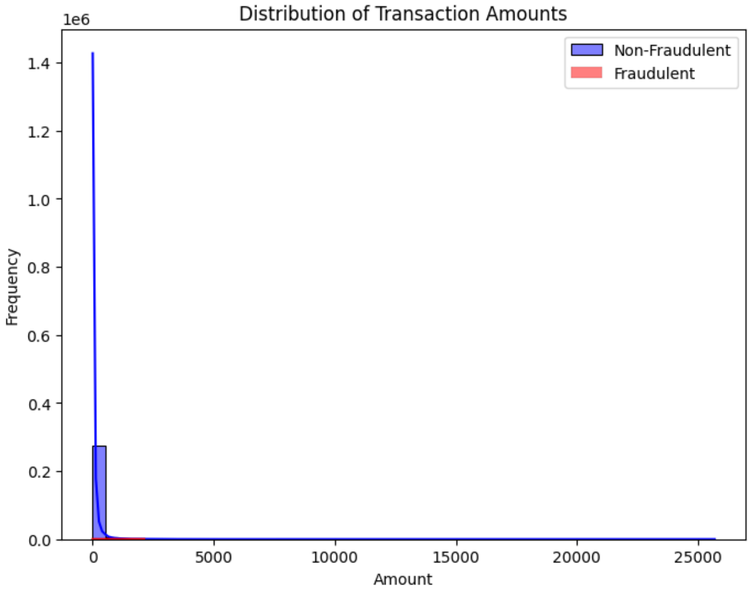

Figure 21: Distribution of Transaction Amounts

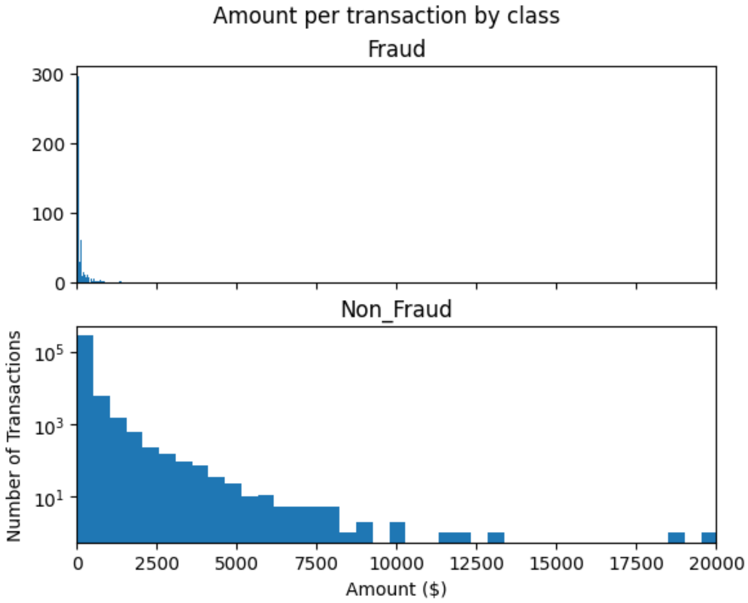

Figure 22: Amount per transactions by class

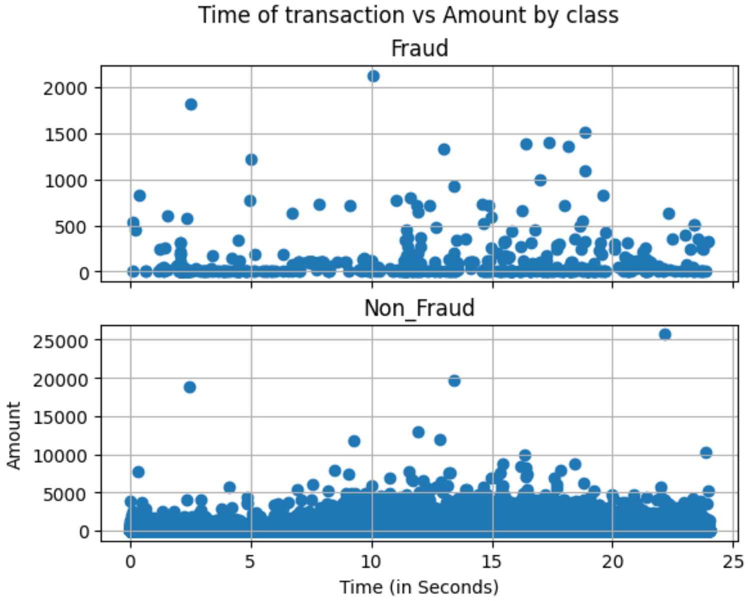

Figure 23: Transaction time versus amount by class

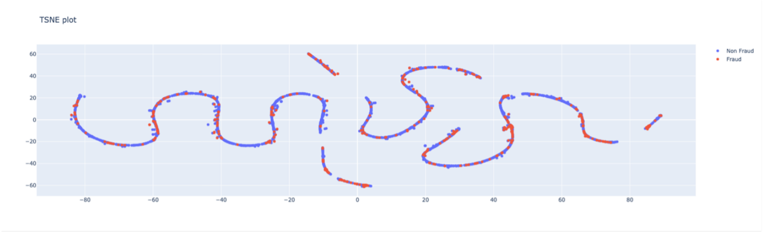

Figure 24: Data visualization with TSNE

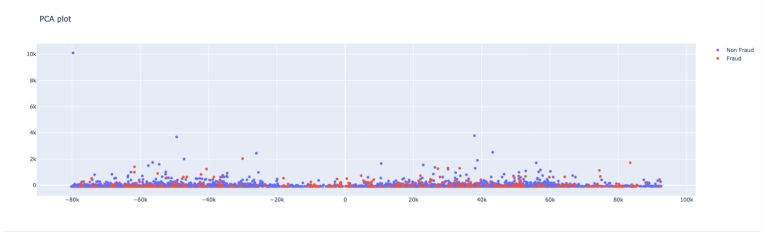

Figure 25: Data visualization with PCA

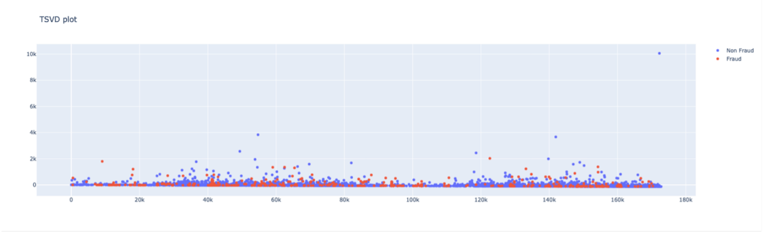

Figure 26: Data visualization with TSVD

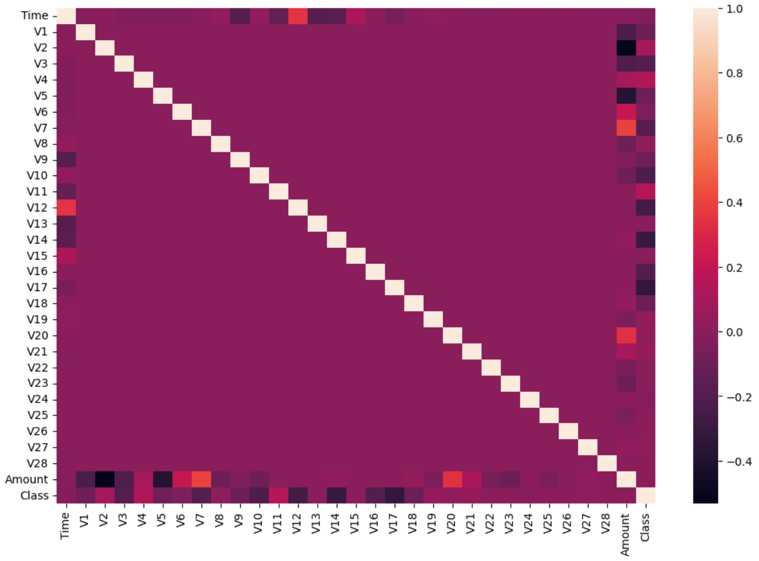

Figure 27: Correlation Matrix

APPENDIX C

Table 7: Software used

Package	Version	Reference
ChatGPT	3.5.0	OpenAI (2024)
Grammarly	April 2024	Grammarly (2024)
Google Colab	April 2024	Google Colaboratory (2024)
Python	3.12.2	Python
Jupyter	7.1.2	(Kluyver et al., 2016)
Pandas	2.2.1	Pandas
NumPy	1.26.4	(Harris et al., 2020)
Matplotlip	3.8.4	Matplotlib
Tensorflow	2.16.1	Tensorflow
Seaborn	0.13.2	(Waskom, 2021)
Scikit-learn	1.4.1	(Pedregosa et al., 2011)
Imbalanced-learn	0.12.2	(Lemaître et al., 2017)
Keras	2.10.0	(Chollet et al., 2015)
Pip	24.0.0	Pip