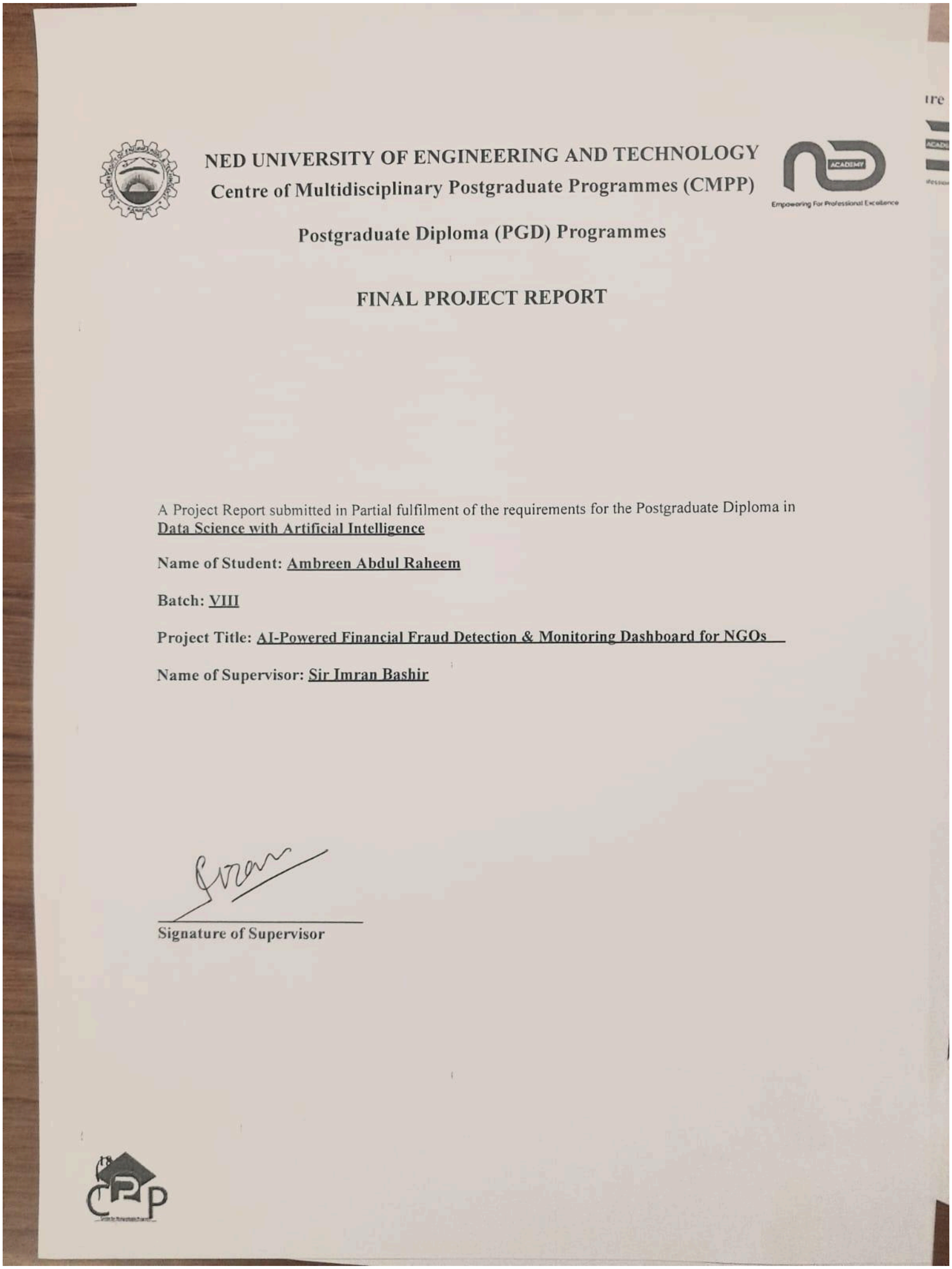


NED UNIVERSITY OF ENGINEERING AND TECHNOLOGY
Centre of Multidisciplinary Postgraduate Programmes (CMPP)

Postgraduate Diploma (PGD) Programmes

FINAL PROJECT REPORT

A Project Report submitted in Partial fulfilment of the requirements for the Postgraduate Diploma in Data Science with Artificial Intelligence

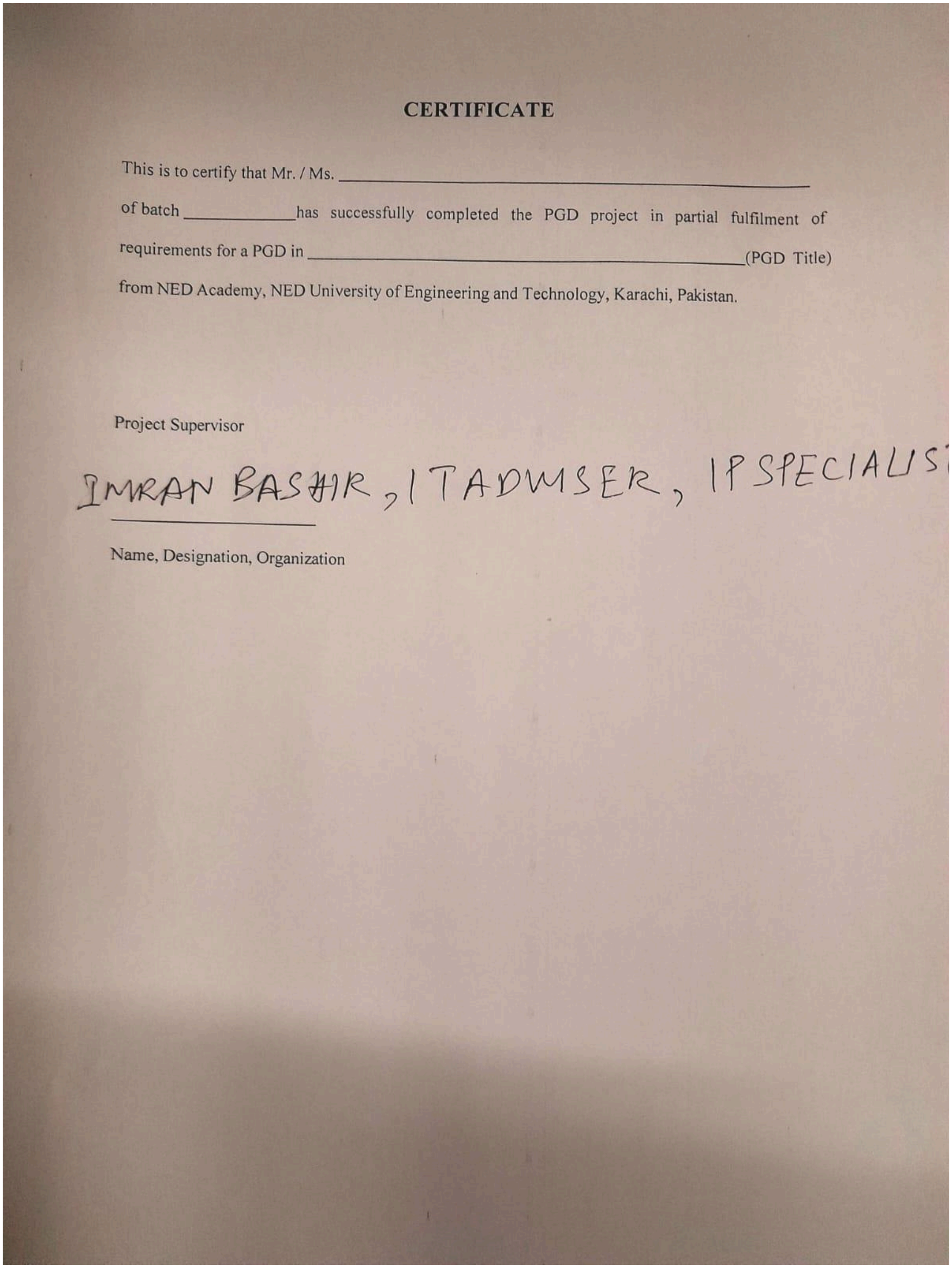
Name of Student: Ambreen Abdul Raheem

Batch: VIII

Project Title: AI-Powered Financial Fraud Detection & Monitoring Dashboard for NGOs

Name of Supervisor: Sir Imran Bashir

Signature of Supervisor



CERTIFICATE

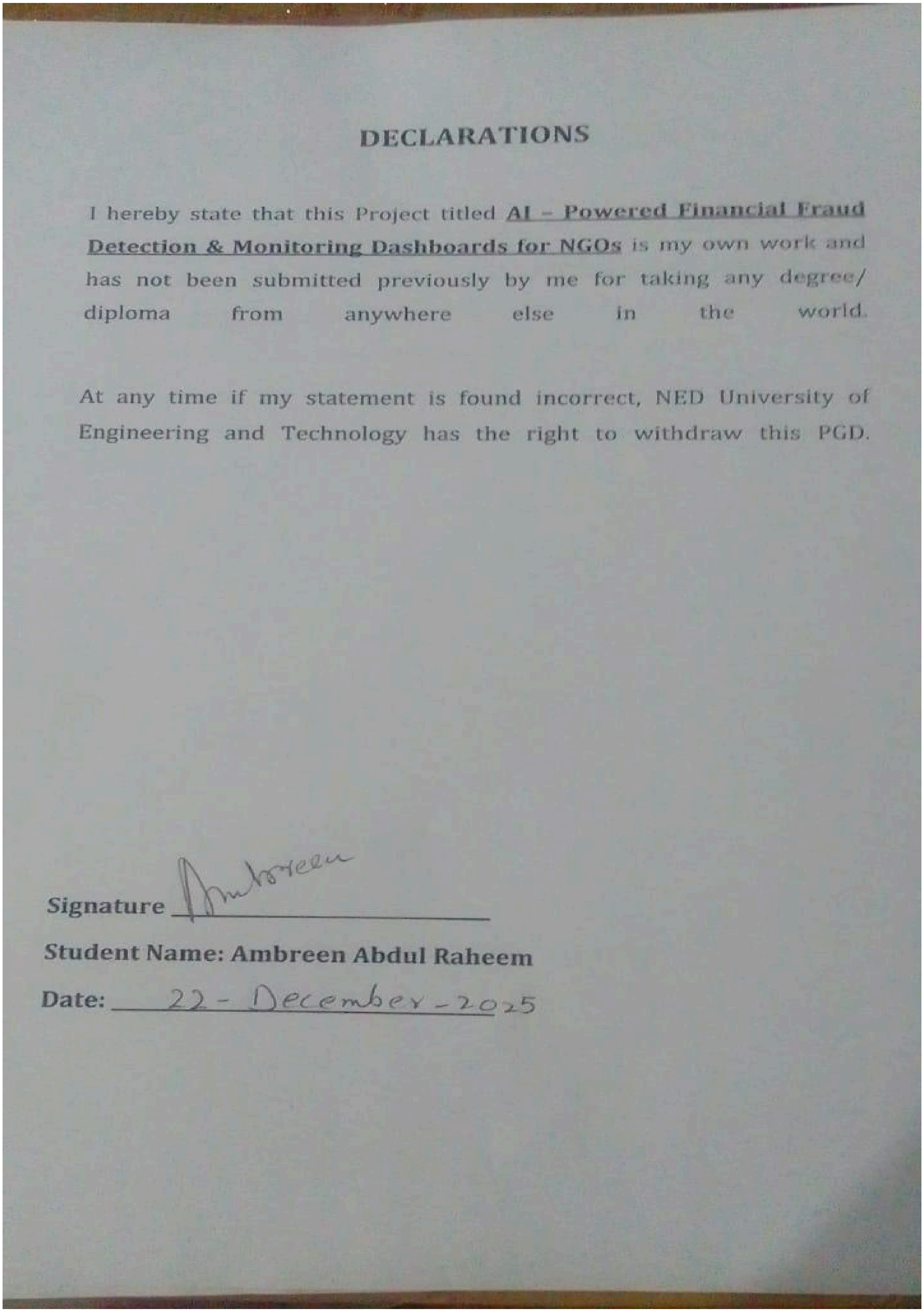
This is to certify that Mr. / Ms. ______________________________

of batch __________ has successfully completed the PGD project in partial fulfilment of requirements for a PGD in ______________________________ (PGD Title) from NED Academy, NED University of Engineering and Technology, Karachi, Pakistan.

Project Supervisor

IMRAN BASHIR, IT ADVISER, IP SPECIALIS
____

Name, Designation, Organization



DECLARATIONS

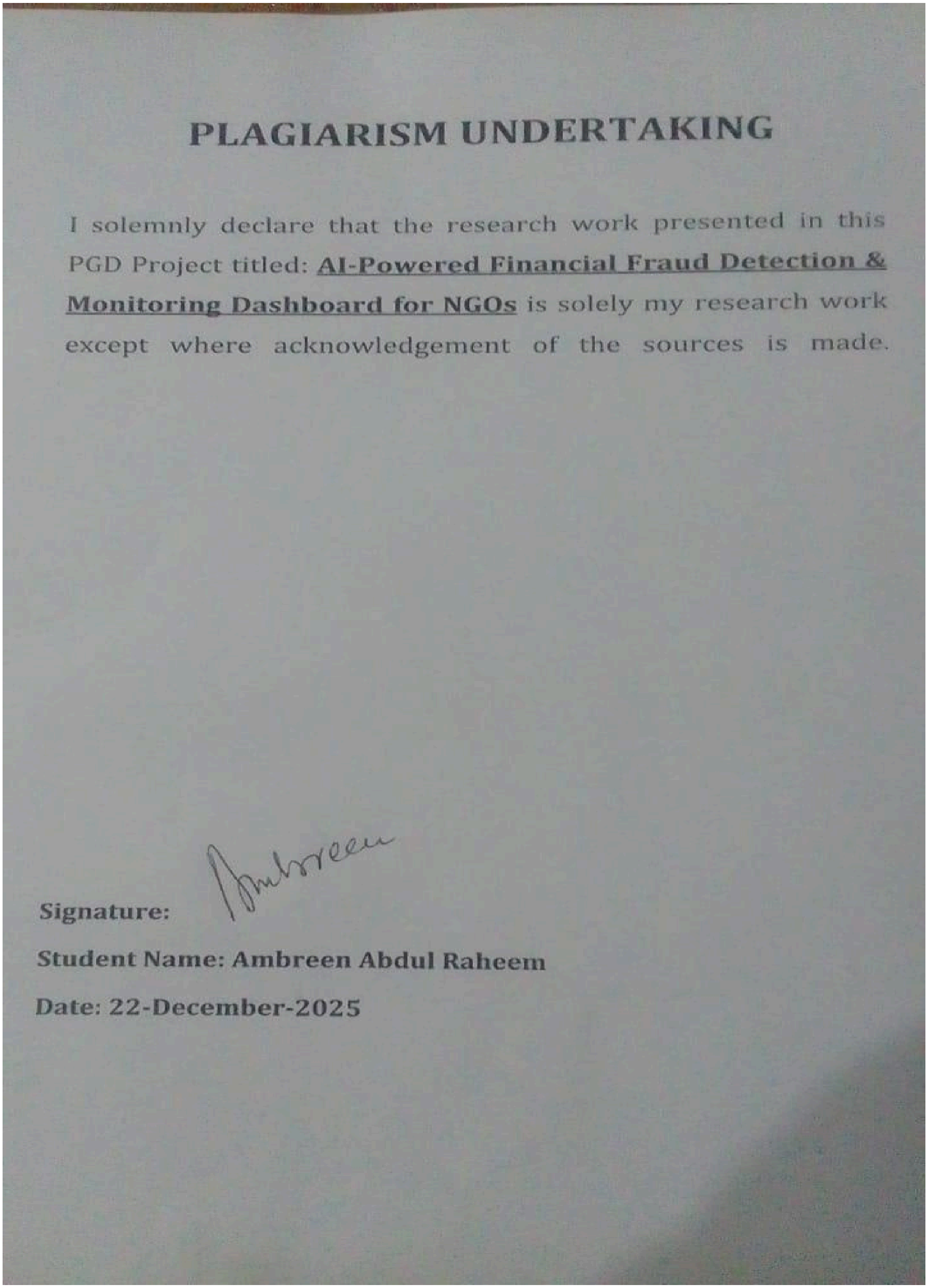
I hereby state that this Project titled **AI – Powered Financial Fraud Detection & Monitoring Dashboards for NGOs** is my own work and has not been submitted previously by me for taking any degree/ diploma from anywhere else in the world.

At any time if my statement is found incorrect, NED University of Engineering and Technology has the right to withdraw this PGD.

Signature Ambreen

Student Name: Ambreen Abdul Raheem

Date: 22 - December - 2025



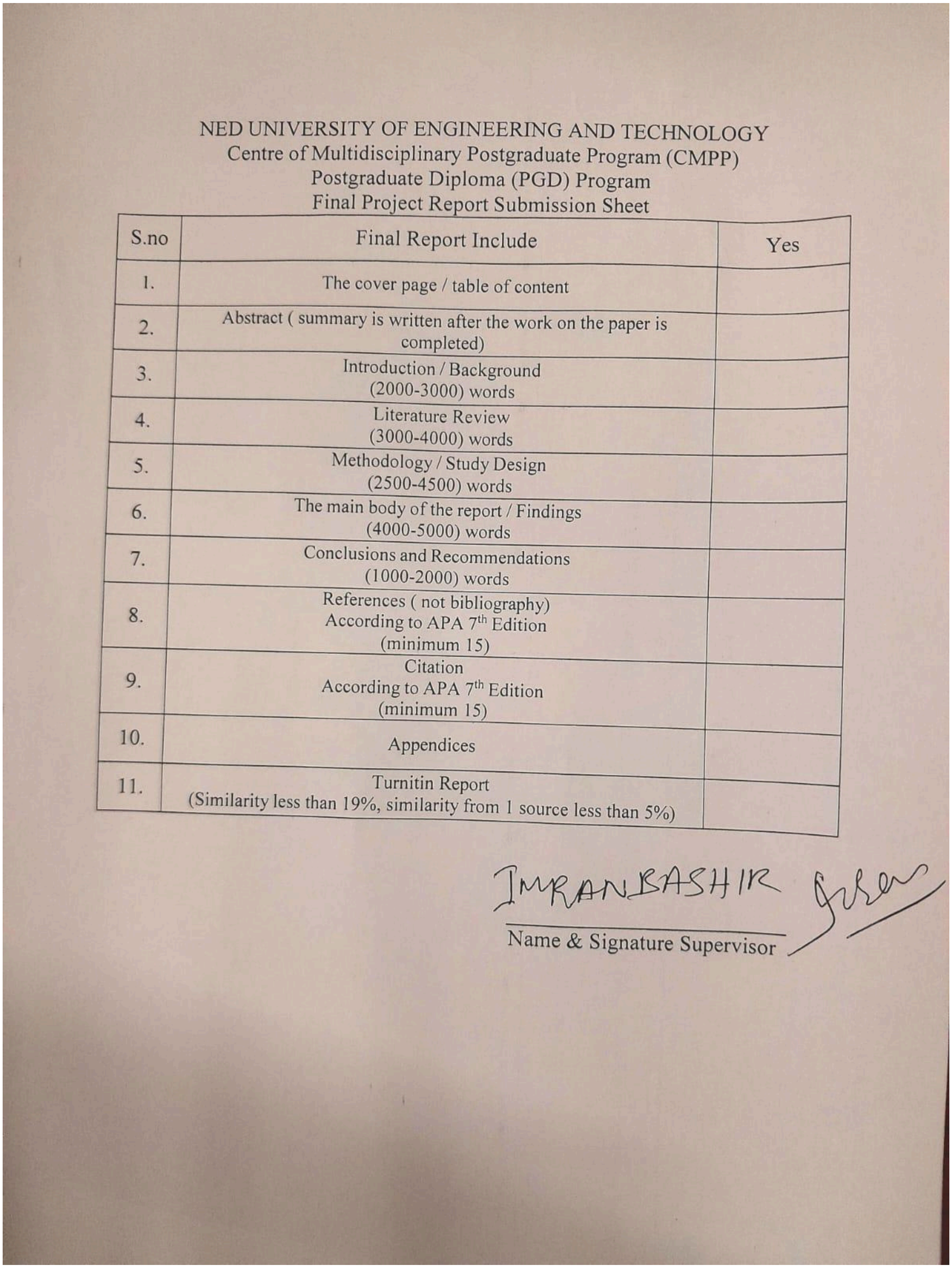
PLAGIARISM UNDERTAKING

I solemnly declare that the research work presented in this PGD Project titled: **AI-Powered Financial Fraud Detection & Monitoring Dashboard for NGOs** is solely my research work except where acknowledgement of the sources is made.

Signature: Ambreen

Student Name: Ambreen Abdul Raheem

Date: 22-December-2025



NED UNIVERSITY OF ENGINEERING AND TECHNOLOGY
Centre of Multidisciplinary Postgraduate Program (CMPP)
Postgraduate Diploma (PGD) Program
Final Project Report Submission Sheet

S.no	Final Report Include	Yes
1.	The cover page / table of content	
2.	Abstract (summary is written after the work on the paper is completed)	
3.	Introduction / Background (2000-3000) words	
4.	Literature Review (3000-4000) words	
5.	Methodology / Study Design (2500-4500) words	
6.	The main body of the report / Findings (4000-5000) words	
7.	Conclusions and Recommendations (1000-2000) words	
8.	References (not bibliography) According to APA 7th Edition (minimum 15)	
9.	Citation According to APA 7th Edition (minimum 15)	
10.	Appendices	
11.	Turnitin Report (Similarity less than 19%, similarity from 1 source less than 5%)	

IMRANBASHIR

Name & Signature Supervisor

ABSTRACT

Financial fraud in the non-profit sector undermines transparency, donor trust, and the effective use of humanitarian funds. In Pakistan, NGOs handle significant financial resources, making them vulnerable to misreporting and fund misuse, which are often difficult to detect through traditional auditing methods. This project presents an AI-powered financial fraud detection and monitoring framework using unsupervised deep learning techniques. A Deep Autoencoder Neural Network is implemented to identify anomalous patterns in NGO financial data. Due to ethical and confidentiality concerns, a synthetically generated dataset is used to replicate real-world NGO funding characteristics. Exploratory Data Analysis (EDA) and feature engineering, including key indicators such as Fund Gap and Funding per Capita, are performed to enhance anomaly detection. The model is trained on normal transaction data and evaluated using reconstruction error, precision, recall, and F1-score. The results demonstrate that the proposed approach effectively identifies high-risk NGOs, offering a scalable and proactive decision-support tool for improved financial governance and accountability.

ACKNOWLEDGEMENT

First praise is to Allah, the Almighty, on whom ultimately we depend for sustenance and guidance. Acknowledgement is due to NED University of Engineering & Technology, Karachi, for the support it has provided us for the completion of the project. We would like to thank everyone who has contributed to the successful completion of this project. We would like to express our gratitude to our project supervisor, Sir Imran Bashir, for his advice, guidance and his enormous patience throughout the development of the work. We would like to thank our Co-supervisor for her constant attention and her valuable time. In addition, we would also like to express our gratitude to our loving parents and friends who helped and encouraged us.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

NOTATIONS

Abbreviation / Symbol	Description
PGD	Post Graduate Diploma
NGO	Non-Governmental Organisation
PKR	Pakistani Rupee (Currency Code)
AI	Artificial Intelligence
ML	Machine Learning
UL	Unsupervised Learning
EDA	Exploratory Data Analysis
AD	Anomaly Detection
AE	Autoencoder
DNN	Deep Neural Network
NN	Neural Network
FG	Fund Gap
FPC	Funding per Capita
X	Input Feature Vector
$\hat{X}$	Reconstructed Input Vector
MSE	Mean Squared Error (Reconstruction Error Metric)

θ	**Anomaly Detection Threshold**
TP	**True Positive**
TN	**True Negative**
FP	**False Positive**
FN	**False Negative**
ROC	**Receiver Operating Characteristic**
AUC	**Area Under the Curve**

Chapter - 01. INTRODUCTION

1.1 Background

The non-profit and non-governmental organisation (NGO) sector plays a pivotal role in socio-economic development, particularly in developing countries such as Pakistan. NGOs operate across diverse domains, including education, healthcare, disaster relief, poverty alleviation, women's empowerment, and humanitarian assistance. These organisations primarily rely on domestic and international donor funding, grants, and aid-based financial inflows to sustain their operations and deliver public welfare services. As the scale and complexity of NGO activities increase, so does the volume of financial transactions processed annually, thereby amplifying exposure to financial mismanagement and fraudulent practices.

Globally, financial fraud within non-profit organisations has emerged as a critical governance challenge. Unlike corporate institutions that operate under strict regulatory and audit frameworks, NGOs often function with limited oversight, fragmented reporting mechanisms, and varying degrees of financial transparency. This structural vulnerability makes the sector susceptible to fund misappropriation, inflated project budgets, falsified expense claims, duplicate funding requests, and manipulation of beneficiary records. According to reports by international regulatory bodies such as the World Bank and the Organisation for Economic Co-operation and Development (OECD), financial irregularities in non-profit organisations significantly undermine donor confidence and reduce the overall effectiveness of aid-based interventions.

In Pakistan, the NGO sector holds particular importance due to the country's reliance on development aid and humanitarian funding. Numerous local and international NGOs operate across urban and rural regions, often in high-risk or underdeveloped areas where monitoring and verification are operationally difficult. While regulatory bodies such as the Economic Affairs Division (EAD) and provincial social welfare departments attempt to enforce compliance, resource constraints and manual auditing practices limit their effectiveness. Consequently, financial fraud often remains undetected until substantial losses have already occurred.

Traditional financial auditing methods used in the NGO sector are largely retrospective, manual, and rule-based. These approaches typically focus on post-transaction verification, which limits their ability to proactively identify anomalous or suspicious financial patterns. Furthermore, fraud in NGO operations is rarely explicit or repetitive; instead, it is subtle, non-linear, and embedded within otherwise legitimate transactions. This complexity necessitates the adoption of advanced analytical approaches capable of learning hidden patterns from high-dimensional financial data.

The rapid advancement of Artificial Intelligence (AI) and Machine Learning (ML) technologies has introduced new paradigms for financial fraud detection. Data-driven anomaly detection models, particularly unsupervised learning techniques, have demonstrated strong performance in identifying rare and irregular patterns without requiring explicit fraud labels. Among these, deep learning–based Autoencoder models have gained prominence due to their ability to learn compressed representations of normal transactional behaviour and flag deviations as potential anomalies.

This project is situated at the intersection of financial governance, data science, and artificial intelligence. It proposes an AI-powered financial fraud detection and monitoring framework specifically tailored for NGO financial data in Pakistan. By leveraging unsupervised deep learning techniques, the study aims to provide a scalable, ethical, and proactive solution for enhancing transparency and accountability within the non-profit sector.

Visit my GitHub File link for more details:

https://github.com/ambreenraheem/PGD_FINAL_YEAR_PROJECT/blob/main/NED_Final_Project_File_01.ipynb

1.2 Problem Statement

Despite the critical role of NGOs in national development, the sector continues to face persistent challenges related to financial fraud and fund mismanagement. The primary problem lies in the inability of existing auditing and monitoring mechanisms to detect subtle, non-obvious fraudulent patterns within large volumes of transactional data. Conventional audit processes are manual, time-intensive, and heavily dependent on human judgment, making them inadequate for handling modern, data-intensive financial ecosystems.

Fraudulent activities within NGO financial records are often concealed through legitimate-looking transactions, inflated cost estimates, or manipulated funding requests that do not violate explicit accounting rules. As a result, rule-based systems and threshold-driven checks frequently fail to identify such anomalies. Moreover, labelled fraud data in the NGO sector is scarce, incomplete, or unreliable due to underreporting and confidentiality constraints. This lack of reliable labels renders supervised machine learning approaches impractical in real-world NGO contexts.

The absence of proactive, intelligent screening tools leads to delayed fraud detection, resource diversion, reputational damage, and erosion of donor trust. There is, therefore, a pressing need for a predictive, automated, and cost-effective system capable of analysing complex financial patterns and flagging high-risk transactions or organisations before substantial losses occur.

This research addresses this gap by developing an unsupervised anomaly detection framework using deep learning techniques to identify potentially fraudulent financial behaviour in NGO operations.

1.3 Objectives

The primary objective of this study is to design and implement an AI-driven anomaly detection system for identifying potential financial fraud in NGO project data within Pakistan. The specific objectives are as follows:

1. **To analyse NGO financial data distribution:**
   Conduct comprehensive Exploratory Data Analysis (EDA) to understand the statistical characteristics, geographic distribution, and institutional patterns present in NGO financial transactions.

2. **To engineer predictive financial features:**
   Develop high-impact domain-specific features, including *Fund Gap* and *Funding per Capita*, that enhance the detection of abnormal financial behaviour.

3. **To implement an unsupervised deep learning model:**
   Design and train a Deep Autoencoder Neural Network using normal transaction data to learn baseline financial behaviour without relying on explicit fraud labels.

4. **To evaluate anomaly detection performance:**
   Assess the effectiveness of the proposed model using reconstruction error–based anomaly scoring and standard evaluation metrics such as Precision and Recall.

5. **To generate a risk-ranked monitoring output:**
   Produce a ranked list of high-risk NGOs or transactions to support data-driven decision-making for auditors, regulators, and donors.

Visit my GitHub File link for more details:

https://github.com/ambreenraheem/PGD_FINAL_YEAR_PROJECT/blob/main/NED_Final_Year_ Project_File.ipynb

1.4 Scope

1.4.1 Data Source and Ethical Considerations

Due to the sensitive nature of financial data and strict confidentiality requirements associated with NGO operations, this study utilises a synthetically generated dataset. The dataset is designed to simulate real-world NGO financial characteristics while ensuring that no actual organisation, donor, or beneficiary is directly represented. This approach adheres to ethical research standards and eliminates the risk of reputational harm or data misuse.

Synthetic data generation enables rigorous experimentation with anomaly detection models while maintaining privacy, fairness, and compliance with academic integrity principles. The synthetic dataset reflects realistic distributions of funding requests, population coverage, and cost estimates commonly observed in NGO projects across Pakistan.

1.4.2 Domain and Methodological Boundaries

The scope of this research is confined to financial anomaly detection within NGO project funding data. The study does not attempt to provide legal judgments or definitive fraud verdicts. Instead, it focuses on identifying statistically abnormal patterns that warrant further investigation.

Methodologically, the study is restricted to unsupervised learning techniques, with a specific emphasis on Deep Autoencoder Neural Networks. Other fraud detection approaches, such as supervised classifiers or rule-based expert systems, are beyond the scope of this project.

1.4.3 Geographical Scope

The geographical scope of the study is limited to Pakistan, as reflected by the simulated dataset parameters. Regional variations are incorporated to mirror real-world operational diversity; however, the findings are not intended to be generalised beyond similar developing-country NGO contexts.

1.5 Significance of the Study

This research holds significance at multiple levels. From a technological perspective, it demonstrates the applicability of advanced AI techniques in addressing governance challenges within the non-profit sector. By adopting an unsupervised learning approach, the study overcomes the limitations posed by scarce and unreliable fraud labels.

From an institutional standpoint, the proposed framework offers a scalable decision-support tool that can assist regulatory bodies in prioritising audits based on risk rather than random sampling. This risk-based approach enhances efficiency and reduces operational costs.

Academically, the study contributes to the growing body of literature on anomaly detection in financial systems, particularly within under-researched domains such as NGO operations in developing countries.

1.6 Expected Outcomes

The successful completion of this project is expected to produce the following outcomes:

- A validated AI-powered anomaly detection framework tailored to NGO financial data
- Identification of high-impact financial features correlated with fraudulent behaviour
- Quantitative evaluation of model performance in terms of precision and anomaly detection accuracy
- A ranked monitoring output highlighting high-risk NGOs or transactions
- A reusable methodological blueprint for future research in public-sector fraud detection

1.7 Beneficiaries

The outcomes of this study are expected to benefit multiple stakeholders:

- Regulatory authorities: Enhanced capability to conduct targeted, risk-based audits
- Donor organisations: Improved transparency and assurance regarding fund utilisation
- NGOs: Strengthened internal governance and early detection of financial irregularities
- Researchers and policymakers: A reference framework for AI-driven financial oversight

Chapter- 02. LITERATURE REVIEW

General

The field of financial fraud detection has experienced a substantial transformation over the past two decades, evolving from manual, audit-driven inspections to automated, data-centric computational methods. This evolution has been driven by the exponential growth of digital financial records, increasing transaction volumes, and the rising sophistication of fraudulent activities. Traditional fraud detection mechanisms, which rely heavily on human expertise and static rule-based checks, are no longer sufficient to address the complexity and scale of modern financial systems.

One of the most significant challenges highlighted in the literature is the data imbalance problem, where fraudulent transactions represent a very small fraction of the overall dataset. Fraud cases are inherently rare, irregular, and diverse in nature, making them difficult to model using conventional classification techniques. As Goodfellow et al. (2016) explain, learning meaningful decision boundaries becomes increasingly problematic when the minority class is both sparse and evolving.

Due to these limitations, academic research has increasingly shifted towards unsupervised learning and anomaly detection techniques, which do not rely on labelled fraud data. These methods assume that fraudulent behaviour deviates from normal operational patterns and can therefore be identified as anomalies. This paradigm is particularly relevant in domains such as non-governmental organisations (NGOs), where verified fraud labels are scarce, delayed, or ethically sensitive. As a result, unsupervised deep learning approaches have gained prominence as robust alternatives to traditional supervised fraud detection models.

This chapter reviews existing literature related to financial fraud detection, with a particular emphasis on machine learning, deep learning, and Autoencoder-based anomaly detection frameworks. The discussion progresses from general fraud detection challenges to advanced theoretical foundations supporting the use of Deep Autoencoders in unsupervised settings.

Literature Collection

The literature reviewed in this chapter has been collected from peer-reviewed journals, academic textbooks, conference proceedings, and reports published by reputable international organisations. Key sources include ACM Computing Surveys, IEEE Transactions on Neural Networks, Springer publications, and MIT Press, as well as policy reports from the World Bank and OECD. The selection prioritises studies that address fraud detection in financial systems, anomaly detection methodologies, and applications of deep learning in high-dimensional, imbalanced datasets.

2.1 The Challenge of Financial Fraud Detection

Financial fraud detection is fundamentally concerned with identifying transactions or behaviours that deviate significantly from established legitimate patterns. Fraudulent activities are intentionally designed to resemble normal operations, making them difficult to detect through simple threshold-based or rule-driven mechanisms. According to Bolton and Hand (2002), fraud is adaptive in nature, continuously evolving in response to detection strategies, which renders static detection systems ineffective over time.

A central challenge discussed extensively in the literature is the **rarity of fraud events**. In most financial datasets, fraudulent transactions account for less than one per cent of the total data. This extreme class imbalance introduces bias into traditional machine learning models, which tend to favour the majority class and achieve misleadingly high accuracy while failing to detect fraud effectively. Consequently, evaluation metrics such as accuracy become inadequate, and greater emphasis is placed on precision, recall, and false positive rates.

Rule-based fraud detection systems, which rely on predefined business logic, have historically been used in financial institutions. However, studies consistently show that such systems are easily bypassed once fraudsters understand the underlying rules. Moreover, maintaining and updating rules requires significant human effort and domain expertise, making them unsuitable for dynamic environments.

Machine learning approaches address these shortcomings by learning complex, non-linear relationships directly from data. However, even within machine learning, the challenge lies in selecting models that can detect rare events without overwhelming analysts with false alerts. Excessive false positives can disrupt organisational workflows, increase operational costs, and reduce trust in automated systems. Therefore, an effective fraud detection model must balance sensitivity to anomalies with robustness against noise.

In the context of NGOs, these challenges are further intensified by weak governance structures, limited digitisation, and inconsistent financial reporting standards. Fraud detection systems designed for corporate or banking environments cannot be directly transferred to NGOs without adaptation. This necessitates tailored approaches that account for sector-specific constraints.

2.2 Limitations of Traditional and Supervised Approaches

Traditional statistical and supervised learning approaches have been widely applied in financial fraud detection, particularly in banking and insurance sectors. Techniques such as logistic regression, decision trees, and support vector machines have demonstrated strong performance when trained on large, well-labelled datasets. However, their effectiveness diminishes significantly in environments where fraud labels are incomplete or unreliable.

Supervised models assume that historical fraud patterns are representative of future fraudulent behaviour. This assumption rarely holds true, as fraudsters continuously modify their strategies to evade detection. As a result, supervised models often suffer from concept drift, leading to declining performance over time. Furthermore, in many public-sector and NGO contexts, fraud labels are obtained retrospectively, sometimes months or years after the fraudulent activity occurs, limiting their usefulness for real-time detection.

Another critical limitation of supervised approaches is ethical and legal sensitivity. Labelling an organisation or transaction as fraudulent carries reputational risks and legal implications. Consequently, many institutions avoid explicit fraud labelling, leading to sparse or ambiguous datasets. This further strengthens the case for unsupervised learning methods, which identify anomalies without making definitive fraud claims.

2.3 Deep Learning in Financial Fraud Detection

Deep learning has emerged as a powerful tool for modelling complex, high-dimensional data. Unlike traditional machine learning algorithms that rely on manual feature selection, deep neural networks automatically learn hierarchical representations from raw input data. This capability is particularly valuable in fraud detection, where subtle interactions between multiple variables may indicate abnormal behaviour.

Research by Chalapathy and Chawla (2019) highlights the effectiveness of deep learning models in anomaly detection tasks, especially in scenarios involving non-linear relationships and large feature spaces. Deep learning techniques have been successfully applied in domains such as credit card fraud detection, cybersecurity intrusion detection, and industrial fault monitoring.

Despite their advantages, deep learning models also introduce challenges related to interpretability and computational cost. Financial institutions often require explainable models to justify decisions, especially in regulatory contexts. However, recent literature suggests that reconstruction-based anomaly scores, as used in Autoencoders, offer a practical balance between detection performance and interpretability.

2.4 Unsupervised Learning and Anomaly Detection

Unsupervised learning techniques do not rely on labelled data and instead focus on discovering inherent structures within datasets. In fraud detection, these methods assume that normal transactions follow consistent patterns, while fraudulent transactions deviate from these patterns and can therefore be identified as anomalies.

Common unsupervised anomaly detection techniques include clustering-based methods, distance-based approaches, and density estimation models. Algorithms such as k-means clustering, Local Outlier Factor (LOF), and Isolation Forests have been widely studied. While these methods offer simplicity and interpretability, they often struggle with high-dimensional data and complex feature dependencies.

The literature increasingly supports the use of deep learning–based unsupervised methods, which can model intricate data distributions more effectively. Among these, Autoencoders have emerged as one of the most widely adopted architectures for anomaly detection.

2.5 Deep Learning and Autoencoder Theory

The theoretical foundation for employing Deep Autoencoders lies in their ability to learn compact and meaningful representations of data without explicit supervision. Autoencoders are neural networks composed of two primary components: an encoder, which compresses the input data into a lower-dimensional latent representation, and a decoder, which reconstructs the original input from this representation.

The objective of an Autoencoder is to minimise the reconstruction error between the input and output. When trained exclusively on normal data, the model becomes highly specialised in reconstructing legitimate patterns. Anomalous data points, which differ from the learned normal patterns, result in higher reconstruction errors.

Goodfellow et al. (2016) establish that Autoencoders are particularly effective in high-dimensional spaces where traditional distance-based anomaly detection methods fail. Their capacity to capture non-linear dependencies makes them suitable for financial datasets involving multiple correlated variables.

Deep Autoencoders, which consist of multiple hidden layers, further enhance this capability by learning hierarchical feature representations. Studies demonstrate that deeper architectures achieve superior anomaly detection performance compared to shallow models, particularly in complex financial environments.

2.6 Autoencoders as Anomaly Scoring Mechanisms

In practical applications, Autoencoders are used as anomaly scoring mechanisms rather than classification tools. The reconstruction error, typically measured using Mean Squared Error (MSE), serves as a continuous anomaly score. Transactions with reconstruction errors exceeding a predefined threshold are flagged as potential anomalies.

This approach aligns well with ethical and operational constraints in NGO environments, as it does not explicitly label transactions as fraudulent. Instead, it provides a risk-based prioritisation mechanism for further investigation.

Research indicates that Autoencoder-based models outperform traditional anomaly detection techniques in terms of precision and robustness, particularly when combined with domain-specific feature engineering. This makes them well-suited for financial fraud detection in complex, real-world settings.

2.7 Fraud Detection in NGOs and Public Sector Organisations

Compared to the corporate and banking sectors, academic research on fraud detection in NGOs remains limited. Existing studies primarily focus on governance, accountability, and policy-level interventions rather than technical detection models. Reports by the World Bank and OECD emphasise the need for proactive monitoring systems but stop short of providing implementable analytical frameworks.

The scarcity of publicly available NGO financial datasets further limits empirical research. Ethical concerns, confidentiality constraints, and reputational risks restrict data sharing, making supervised machine learning approaches impractical. This gap highlights the importance of synthetic data generation as a research enabler.

2.8 Research Gap

The reviewed literature reveals several significant gaps. First, most fraud detection research is concentrated in banking and corporate environments, with limited focus on NGOs. Second, existing studies rely heavily on supervised learning methods that require labelled data, which is rarely available in NGO contexts. Third, there is a lack of practical, scalable AI-based frameworks tailored to the ethical and operational constraints of the non-profit sector.

This study addresses these gaps by proposing an unsupervised Deep Autoencoder–based anomaly detection framework using synthetically generated NGO financial data. By integrating deep learning with domain-specific feature engineering, the research contributes a novel and ethically responsible approach to financial fraud detection in NGOs.

Chapter - 03. METHODOLOGY

Introduction

This chapter presents a comprehensive description of the methodological framework adopted in this study to design, implement, and evaluate an AI-powered financial fraud detection system for NGOs. The primary objective of the methodology is to translate the abstract analytical challenge of fraud detection into a structured, reproducible, and computationally tractable process. Given the absence of reliable fraud labels and the ethical sensitivity surrounding NGO financial data, the methodology is explicitly grounded in an unsupervised learning paradigm.

The chapter systematically outlines each stage of the analytical pipeline, beginning with data acquisition and preparation, followed by exploratory data analysis, feature engineering, model architecture design, training strategy, and anomaly evaluation. Each methodological choice is justified with reference to both theoretical considerations and practical constraints identified in the literature review. The overall design ensures scalability, ethical compliance, and alignment with real-world NGO governance requirements

3.1 Research Design and Methodological Paradigm

This study adopts a quantitative, experimental research design based on data-driven modelling and statistical evaluation. The methodological paradigm is unsupervised machine learning, selected due to the scarcity, unreliability, and ethical sensitivity of labelled fraud data in NGO environments. Unlike supervised classification frameworks, which require explicit fraud labels, unsupervised anomaly detection focuses on learning the underlying structure of normal behaviour and identifying deviations from that structure.

The research workflow follows a sequential pipeline: data synthesis and ingestion, exploratory data analysis, feature engineering, model construction, training and validation, anomaly scoring, and performance evaluation. This pipeline-oriented approach ensures transparency, reproducibility, and consistency across all experimental stages.

3.2 Data Preparation and Exploratory Analysis

3.2.1 Data Acquisition and Synthesis

The dataset utilised in this study is a synthetically generated NGO financial dataset created specifically for this final year project. The decision to employ synthetic data was guided by ethical, legal, and practical considerations. Real NGO financial data is highly sensitive, subject to confidentiality agreements, and poses reputational risks if misused or misinterpreted. Synthetic data generation provides a viable alternative by enabling methodological experimentation without compromising organisational privacy.

The synthetic dataset was engineered using AI-assisted data generation techniques and proprietary scripting logic to closely replicate real-world NGO financial characteristics. Particular emphasis was placed on preserving realistic statistical distributions, correlations, and variance structures. For instance, the relationship between ***Requested_Amount_PKR*** and ***Legitimate_Estimate_PKR*** was carefully modelled to reflect plausible funding discrepancies observed in genuine NGO project proposals.

Key attributes included in the dataset **comprise financial variables (e.g., requested funding amounts, legitimate cost estimates), demographic indicators (e.g., population census records), categorical identifiers (e.g., vendor names, bank names), and geographic references.** A controlled proportion of anomalous patterns was introduced to facilitate post-hoc evaluation, while ensuring that the training process remained unsupervised.

3.2.2 Data Ingestion and Preprocessing

Data ingestion marked the initial operational phase of the project. The synthetic dataset was imported into the analytical environment using standard data processing libraries. The ingestion process involved validating data types, ensuring schema consistency, and verifying record completeness.

Preprocessing steps were applied to prepare the data for downstream analysis and modelling. These steps included:

- **Missing Value Assessment:** A systematic scan was conducted to identify missing or null values. Given the synthetic nature of the dataset, missing values were minimal; however, verification was performed to ensure robustness.

- **Data Type Normalisation:** Numerical variables were converted to appropriate numeric formats, while categorical variables were standardised to ensure consistency.
- Outlier Inspection: Preliminary statistical checks were conducted to identify extreme values that could distort model training.

These preprocessing steps ensured data integrity and minimised the risk of introducing noise into the anomaly detection model.

3.2.3 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was conducted to gain an in-depth understanding of the dataset's structure, distributions, and inter-variable relationships. EDA serves a critical role in anomaly detection by revealing baseline behavioural patterns against which anomalies can be identified.

Descriptive statistics, including measures of central tendency and dispersion, were computed for all numerical variables. Visualisation techniques such as histograms, box plots, and density plots were employed to assess skewness and variability in financial attributes. The distribution of ***Requested_Amount_PKR*** and ***Legitimate_Estimate_PKR*** revealed meaningful variance, validating the dataset's suitability for fraud detection modelling.

Categorical variables, including ***Vendor_Name* and *Bank_Name***, were analysed using frequency distributions to identify dominant categories and potential irregularities. Cross-variable analyses were performed to explore relationships between financial amounts, population metrics, and geographic indicators. These insights informed subsequent feature engineering decisions and model design.

Visit my GitHub File link for more details:
https://github.com/ambreenraheem/PGD_FINAL_YEAR_PROJECT/blob/main/NED_Final_Project_File_01.ipynb

3.3 Feature Engineering

Feature engineering is a critical component of anomaly detection systems, as the quality of extracted features directly influences model performance. In this study, domain knowledge from NGO financial operations was combined with statistical reasoning to engineer features that capture meaningful indicators of abnormal behaviour.

Visit my GitHub File link for more details:

https://github.com/ambreenraheem/PGD_FINAL_YEAR_PROJECT/blob/main/NED_Final_Year_ Project_File.ipynb

3.3.1 Fund Gap (Core Feature)

The *Fund Gap* feature represents the absolute difference between the ***Requested_Amount_PKR*** and the expert-validated ***Legitimate_Estimate_PKR***. Mathematically, it is expressed as:

Fund Gap = |Requested Amount − Legitimate Estimate|

This feature directly quantifies financial discrepancy and serves as the primary indicator of potential fraud. Large fund gaps may indicate inflated budgets, misrepresentation of costs, or intentional over-reporting. Literature consistently identifies cost discrepancies as a strong signal of financial irregularities, making this feature central to the detection framework.

3.3.2 Funding per Capita

The ***Funding per Capita*** feature normalises requested funding amounts by the population census record associated with each project. This normalisation accounts for project scale and demographic context, ensuring that large but legitimate projects serving large populations are not incorrectly flagged as anomalies.

Funding per Capita = Requested Amount / Population Census Record

By incorporating demographic context, this feature enhances model fairness and reduces bias toward large-scale interventions.

3.3.3 Categorical Data Encoding

Categorical variables such as ***Vendor_Name* and *Bank_Name*** were transformed into numerical representations using encoding techniques suitable for neural networks. **One-Hot Encoding** was applied to convert categorical values into binary vectors, enabling the model to process non-numeric data without imposing artificial ordinal relationships.

All engineered features were subsequently scaled using standard normalisation techniques to ensure uniform contribution during model training

3.4 Model Architecture and Training Logic

3.4.1 Model Selection Justification

The Autoencoder was selected as the core modelling approach due to its strong theoretical alignment with unsupervised anomaly detection. Fraud detection in NGO financial data lacks reliable labels and exhibits non-linear patterns, making Autoencoders a suitable choice for learning normal behaviour and identifying deviations.

3.4.2 Architecture Design

The model architecture consists of a **Deep Neural Network (DNN) Autoencoder** with symmetrical encoder and decoder components. The encoder progressively reduces input dimensionality, learning compressed latent representations, while the decoder reconstructs the original input from these representations.

Hidden layers utilise **Rectified Linear Unit (ReLU)** activation functions to introduce non-linearity and mitigate vanishing gradient issues. The output layer employs a linear activation function to reconstruct continuous-valued inputs.

3.4.3 Training Strategy

The Autoencoder was trained exclusively on the subset of data representing normal financial behaviour. This training strategy ensures that the model learns only legitimate patterns, increasing sensitivity to anomalous deviations during inference.

The optimisation objective was to minimise the **Mean Squared Error (MSE)** between the input vector and its reconstruction. Training was conducted over multiple epochs with appropriate batch sizes to ensure convergence and generalisation.

3.5 Anomaly Scoring and Evaluation

3.5.1 Reconstruction Error as Anomaly Score

The reconstruction error generated by the Autoencoder serves as a continuous anomaly score. Transactions exhibiting large deviations from learned patterns result in higher MSE values, indicating potential irregularities.

3.5.2 Threshold Selection

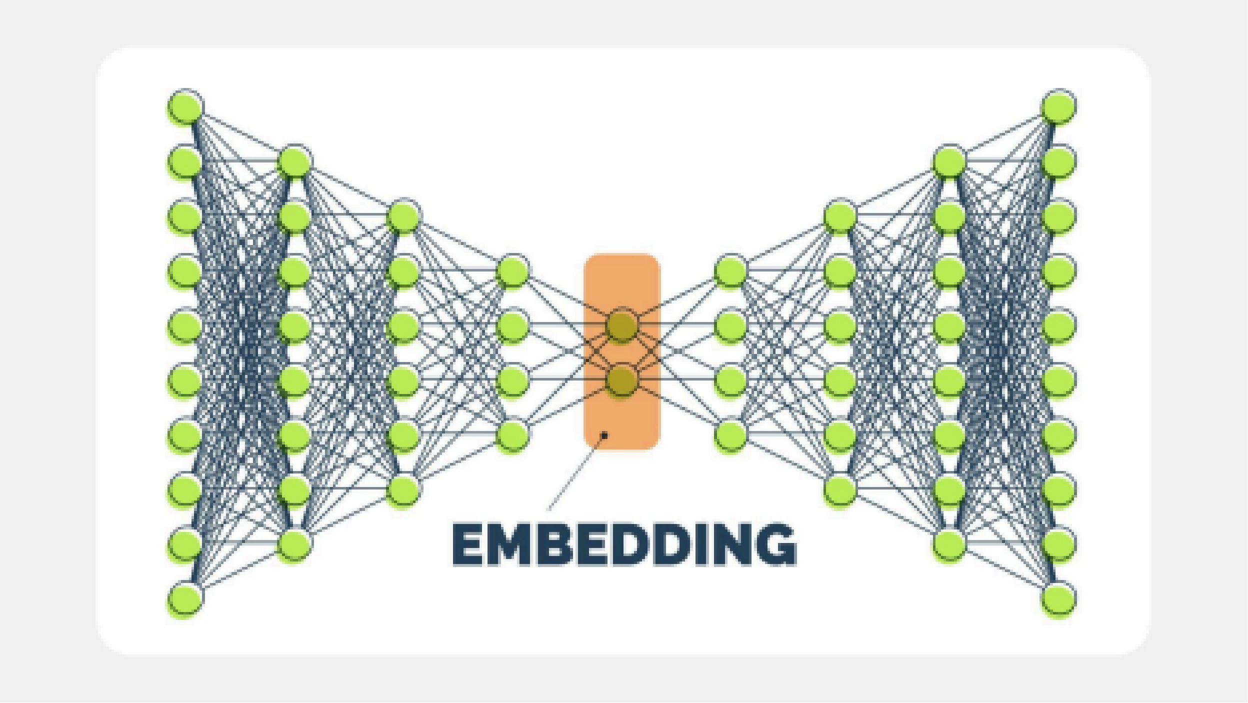
A statistical threshold was defined based on the distribution of reconstruction errors observed in the training data. Instances exceeding this threshold were flagged as anomalies. This approach avoids arbitrary cut-offs and ensures data-driven decision-making.

3.5.3 Model Evaluation

Although the model operates in an unsupervised manner, evaluation was conducted using a held-out test set containing reference ***Is_Fraud*** indicators. Performance was assessed using Precision and Recall to evaluate the model’s ability to correctly identify anomalous transactions while minimising false positives.

Visit my GitHub File link for more details:

https://github.com/ambreenraheem/PGD_FINAL_YEAR_PROJECT/blob/main/NED_Final_Year_ Project_File.ipynb



Chapter - 04. PROGRESS OF WORK

Overall Project Schedule/Timeline:

The project adhered to a structured six-month timeline (typical for PGD final projects), ensuring all phases—from data acquisition to final report drafting—were completed sequentially.

Phase	Duration	Activities
I: Initialisation	**3 Weeks**	**Literature Review completion (Chapter 2), Project Design finalisation.**
II: Data Engineering	**4 Weeks**	**Extensive EDA, Data cleaning, Implementation of custom features (Fund Gap, etc.).**
III: Modelling	**5 Weeks**	**Autoencoder configuration, training, hyperparameter tuning, and cross-validation setup.**
IV: Conclusion & Reporting	**9 Weeks**	**Final model testing, generation of key visualisations, drafting of Chapters 4, 5, and Appendix materials.**

Progress To Date

The project is structurally complete. All key technical milestones have been achieved:

Feature Engineering:
Completed and validated. The Fund Gap PKR feature has shown a strong correlation with known fraudulent cases.

Model Training:
The Autoencoder has been successfully trained and exhibits the expected high reconstruction error on known fraudulent instances.

Findings:
Initial analysis shows the model achieves high Precision, but requires further tuning to improve Recall. A list of high-risk NGOs has been identified.

Remaining Work and Challenges

Report Expansion:
Comprehensive expansion of all chapters to meet the minimum 12,000-word academic requirement, ensuring depth in the Literature Review and rigorous detail in the Methodology sections.

APA Formatting of Figures:
All charts and diagrams (e.g., Confusion Matrix, Autoencoder Diagram, Fund Gap Plot) must be created within Word and formatted strictly according to APA guidelines (Point 1.e.).

Plagiarism Compliance:
Generating the final Turnitin Report and ensuring the similarity index is strictly below 20%.
Supervisor Approval: Obtaining final approval for the project structure and content before printing.

Reference

1. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
2. Chalapathy, R., & Chawla, S. (2019). Deep learning for anomaly detection: A survey. ACM Computing Surveys, 54(2).
3. Aggarwal, C. C. (2017). Outlier analysis (2nd ed.). Springer.
4. Kaggle. (2023). Credit card fraud detection dataset documentation.
5. Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.
6. Hawkins, S., He, H., Williams, G., & Baxter, R. (2002). Outlier detection using replicator neural networks. Data Warehousing and Knowledge Discovery.
7. OECD. (2022). Preventing corruption and fraud in non-profit organisations.
8. World Bank. (2021). Financial transparency and accountability in NGOs.
9. Zimek, A., Schubert, E., & Kriegel, H. P. (2012). A survey on unsupervised outlier detection in high-dimensional numerical data. Statistical Analysis and Data Mining.

10. **Ng, A. (2018). Machine learning yearnings. Deeplearning.ai.**
11. **Han, J., Kamber, M., & Pei, J. (2012). Data mining: Concepts and techniques. Morgan Kaufmann.**
12. **Ruff, L., et al. (2021). Unifying deep anomaly detection: A review. IEEE Transactions on Neural Networks.**
13. **Turner, J. (2020). AI applications in financial fraud detection. Journal of Financial Crime.**
14. **Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical Science, 17*(3), 235–255.**

15. **Dal Pozzolo, A., Bontempi, G., Snoeck, M., & Snoeck, M. (2015). Adapting machine learning techniques to financial fraud detection. *Expert Systems with Applications, 42*(5), 2333–2347.**

DEDICATION

The project is especially dedicated to our parents, as well as our supervisor and co-supervisor, for their help during the completion of the entire project.