UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

Evaluating the Stability of Causal Machine Learning Against Data and Concept Drift in Fraud Detection Systems

Louandra Arjunan

(713982)

School of Mechanical, Aeronautical and Industrial Engineering

University of the Witwatersrand

Johannesburg, South Africa.

Supervisors:

Doctor Bevan Smith, Professor Daniel Wilke

A research **proposal** submitted to the Faculty of Engineering and the Built Environment, University of the Witwatersrand, in fulfilment of the requirements for the degree of Master's of Science in Engineering.

July 2025

Abstract

Fraud causes substantial financial losses and affects customers through evolving attack types such as SIM swap and subscription fraud. Traditional machine learning models used for fraud detection are largely correlational and often fail under dynamic, adversarial conditions. As fraudsters employ novel methods to commit fraud, both data drift (covariate shift) and concept drift are introduced into fraud detection systems, leading to rapid model performance degradation. This research investigates whether causal models provide greater robustness than correlational models under drift. A synthetic data framework will be developed using a Structural Causal Model (SCM) informed by domain knowledge of telecommunications fraud to serve as ground truth. Controlled simulations will introduce abrupt, gradual, and recurring patterns of both data and concept drift. State-of-the-art correlational models (XGBoost, Random Forests, Neural Networks) and causal discovery and inference methods (PC, GES, FCI, LiNGAM) will be evaluated. Model degradation will be assessed using predictive metrics (AUC, precision, recall) and causal stability measures including Structural Hamming Distance (SHD) and Structural Intervention Distance (SID).

Contents

1 RESEARCH CONTEXT **1**

1.1 Background 1

1.2 Motivation 1

1.3 Identified Gaps in Fraud detection 2

1.4 Research Question 3

1.5 Objectives 3

1.6 Delimitations 3

2 LITERATURE REVIEW **4**

2.1 Machine Learning in Fraud Detection 4

2.2 Structural Causal Models (SCMs) 5

2.3 Causal Discovery Methods 5

2.4 Probabilistic Causal Models 6

2.5 Causal Inference and Graph Neural Networks 7

2.6 Drift in Machine Learning 7

2.7 Drift detection techniques 8

3 RESEARCH METHODS **10**

3.1 Synthetic Data Generation and Drift Simulation 10

3.2 Drift Simulation 11

3.3 Baseline Modelling 13

3.4 Iterative Drift Application and Evaluation 13

3.5 Tools and Libraries 14

4 ETHICAL ISSUES **15**

5 PROJECT SCHEDULE **16**

6 REFERENCES **17**

7 Appendices **22**

1 RESEARCH CONTEXT

1.1 Background

In 2023, 15.5 percent of all suspected digital fraud targeted the telecommunications industry, resulting in losses worldwide of $38.95 billion in 2023 (Transunion, 2024) and severe reputational damage. Types of fraud impacting telecommunications include subscription fraud, which broadly describes any type of fraud where a fraudster creates an account with services for which they do not intend to pay (Babaei, et al., 2019). Subscription fraud is either committed for profit (such as selling the services or devices obtained) or for personal use. It is not easily distinguishable from bad debt (Estevez, et al., 2006). Subscription fraud often results from fraudsters gaining access to just enough customer information through phishing, vishing or smishing to be able to gain control of the account and add new lines to the account for which the customer is charged and the company has to reimburse. SIM swap fraud is another common type of fraud affecting telcos in Africa. A fraudster initiates the transfer of a customer's phone number (MSISDN) and account onto a new SIM card that they possess. This enables them to access cellphone banking and other accounts verified through OTPs (One Time Pins) sent to the cellphone number (Ekeh, et al., 2022). Current methods of prevention and detection in telcos include prediction models trained on prior device or MSISDN activity, and authentication questions to verify that the requester is the owner of the MSISDN. The answers to these authentication questions can sometimes be easily phished by the fraudster, allowing them to proceed with swaps. The availability of stolen credentials and personal information through increasingly common corporate data breaches makes it easier for fraudsters to bypass knowledge-based authentication and gain control of a victim's account. The rise of generative and agentic AI has further empowered fraudsters in rapidly adapting their techniques to outsmart existing systems. Malicious AI agents can use hyper-personalized social engineering and deepfakes for phishing and using synthetic identities to fraudulently acquire services. Autonomous agents can continuously probe systems for vulnerabilities, creating a more persistent and adaptive threat than traditional fraud methods (Fenstermacher, 2025).

1.2 Motivation

Organisations require robust fraud detection systems to prevent financial and reputational losses. Current machine learning and data driven fraud detection and prevention methods in telecommunications companies are based on associational relationships and often do not capture the causal mechanisms driving fraud. There is also a need for enhanced explainability of machine learning

models that are customer-facing. Drift poses the largest challenge to the reliability of fraud detection systems as models need to operate in dynamic environments. Machine learning models learn correlations from historical data, the accuracy of which can significantly degrade over time due to evolving fraud tactics (Webb, et al., 2025), leading to costly increases in undetected fraud or a rise in false alarms leading to customer dissatisfaction. Current approaches for addressing model drift in fraud detection models include reactive strategies, such as retraining the model periodically when new information surfaces, and sophisticated drift detection methods which may only address the symptoms rather than the root cause of the performance decay. The motivation for this research is the hypothesis that causal models offer more inherent robustness to model drift than statistical models. By learning the invariant mechanisms that drive fraudulent activity rather than just correlations, causal models theoretically should be less susceptible to the performance degradation caused by shifts in data distributions or even changes in the underlying causal pathways of fraud.

1.3 Identified Gaps in Fraud detection

- Identifying novel methods used to commit fraud: Models are typically created to detect particular methods of fraud. Novel techniques to bypass these detection methods are devised by fraudsters at speed and scale and the lag between implementation and detection results in significant financial loss and customer experience failures.

- Explainability: Many machine learning models used to build fraud detection systems lack the interpretability required by business stakeholders to adopt them in customer facing environments due to the customer service risk. Current explainability frameworks explain correlations (such as feature importance via SHAP), not causation, risking misplaced confidence in spurious relationships.

- Limited understanding of causal model behaviour under drift: Most drift adaptation research focuses on traditional machine learning, with little exploration of how structural causal models degrade over time.

- Lack of comparative studies between causal and associational models: While some works study drift adaptation, few directly compare causal models with non-causal models under controlled drift conditions.

- Lack of established framework for evaluating causal model drift in fraud detection: Traditional drift detection may not be effective for causal models in production settings.

1.4 Research Question

How robust are causal machine learning models compared to traditional statistical models in sustaining predictive performance under controlled data drift and concept drift?

1.5 Objectives

1. Create a software package for the generation of synthetic datasets with known causal structures and the simulation of data and concept drift.

2. Establish baseline predictive performance and initial causal-effect estimation accuracy for correlational and causal models on the non-drifted synthetic dataset through training and evaluating both correlational models alongside causal models.

3. Quantify and compare the sustained predictive performance of traditional and causal models across all simulated drift iterations. Measure and compare model performance degradation (changes in machine learning model metrics), feature-drift metrics, and causal-graph changes.

4. Assess and compare the stability and accuracy of causal models against the known ground truth across all drift iterations to determine robustness.

5. Translate the experimental findings on model robustness into recommendations for causal machine learning architectures suitable for interpretable and effective SIM-swap and subscription fraud detection.

1.6 Delimitations

- The research will use synthetic datasets with predefined causal structures.

- Real-world datasets will be used to test the approaches designed on the synthetic datasets.

2 LITERATURE REVIEW

2.1 Machine Learning in Fraud Detection

Telecommunications customers have been attractive targets for fraudsters due to the difficulty companies face in detecting and localising the source, the inexpensive means of committing fraud on a network, and the various entry points (Hilas, 2009). Rule-based fraud detection systems are characterised by a set of rules defined by a subject matter expert with prior knowledge of fraud scenarios (Babaei et al., 2020). While these systems are straightforward and interpretable, a major drawback is that fraudsters constantly adapt and update their methods to exploit vulnerabilities, and rule-based systems are ineffective at detecting new fraud patterns in a timely manner (Nobel et al., 2024). Machine learning algorithms are able to adapt to changing fraudulent behaviour and uncover intricate patterns in data to a larger extent (Wang et al., 2025). Supervised machine learning algorithms require a labelled set of past events to learn a prediction model. The label (fraudulent or not fraudulent) is usually determined some time after the event has occurred, following investigation (Carcillo et al., 2019). Unsupervised methods aim to characterise the distribution of the data, relying on the assumption that outliers indicate fraud. These algorithms do not require labelled data and are useful for detecting previously unseen fraud types. Unsupervised techniques include clustering and dimensionality reduction (Carcillo et al., 2019). Table 1 summarises commonly used supervised learning models for fraud detection in various studies.

Table 1: Supervised Methods for Fraud Detection

Algorithm	**Dataset**	**Performance**	**Reference(s)**
Logistic Regression	IEEE CIS Fraud Detection	• Precision 0.95 • Recall 0.96 • F1 score 0.95 • AUC 0.82	(Wang et al., 2025)
Random Forest	IEEE CIS Fraud Detection, European Credit Card Dataset	• Precision 0.97 - 0.98 • Recall 0.86 - 0.97 • F1 score 0.91 - 0.97 • AUC 0.91	(Wang et al., 2025), (Wijaya et al., 2024)

LightGBM	IEEE CIS Fraud Detection	• Precision 0.97 • Recall 0.97 • F1 score 0.97 • AUC 0.90	(Wang et al., 2025)
XGBoost (Extreme Gradient Boosting)	European Credit Card Dataset	• Precision 0.97 • Recall 0.88 • F1 score 0.92	(Wijaya et al.,2024)

2.2 Structural Causal Models (SCMs)

Structural Causal Models (SCMs) are a framework for representing causal relationships using directed graphical models (Pearl, 2000). SCMs can be used to model how variables in a system influence each other. Unlike traditional correlational models that rely on associations, SCMs enable explicit modelling of the data-generating process and provide tools for answering counterfactual queries. Two components make up the SCM, a Directed Acyclic Graph (DAG) and a set of structural equations which expresses each variable's value as a function of its parent nodes or unobserved noise (Neal, 2020).

2.3 Causal Discovery Methods

Causal discovery is the process of learning causal relationships from observational data. Table 2 summarises widely used causal discovery methods.

Table 2: Causal Discovery Methods

Name	**Short Description**	**Strengths**	**Weaknesses**	**References**
PC	Uses conditional independence (CI) tests to infer causal structure assuming causal sufficiency and faithfulness.	Simple, well-established, efficient for small/medium graphs.	Sensitive to CI test errors; cannot handle latent confounders.	Neal (2020)

FCI	Extension of PC that infers structure under latent confounding and selection bias. Outputs Partial Ancestral Graphs (PAGs).	Handles hidden confounders, uses conservative edge orientation and is flexible and non-parametric. RFCI, FCI+ and GFCI are improvements on FCI.	Computationally intensive, output can include ambiguous edges (circles) and CI tests can be unreliable with noise or small samples.	Spirtes et al. (2000), Colombo et al. (2012), Claassen et al. (2013), Ogarrio et al. (2016)
GES	Score-based algorithm optimizing a score (such as BIC) to find causal structure. Outputs a PDAG.	Uses a systematic scoring approach with flexible structure updates.	Requires causal sufficiency and may get stuck in local optima.	Chickering (2002)
LiNGAM	Assumes linear, non-Gaussian models with no unobserved confounders. Uses non-Gaussianity to infer directionality.	Can identify full causal structure under its assumptions.	Requires linearity and no hidden confounders.	Shimizu et al. (2006)
NOTEARS	Optimization-based approach using continuous constraints and gradient descent for structure learning.	Highly scalable and efficient for high-dimensional data.	Assumes linearity and may miss complex non-linear causal relations.	Zheng et al. (2018)

2.4 Probabilistic Causal Models

A Bayesian Network (BN) is a probabilistic graphical model that represents a set of variables and their conditional dependencies via a DAG. Each node is associated with a conditional probability distribution quantifying the relationship between the node and its direct causes (Moso & Kenei, 2018). A key strength of causal BNs is the ability to integrate both data-driven evidence and expert

domain knowledge (Hu, et al., 2023). The structure of the graph can be defined by fraud analysts, while the conditional probabilities can be learned from historical data. Traditional BNs often require variables to be discretised, which can lead to a loss of information (Mukhanov & Lev, 2008).

2.5 Causal Inference and Graph Neural Networks

Graph Neural Networks (GNNs) have emerged as the state-of-the-art for learning from such graph-structured fraud data, as fraud are inherently network phenomena (Cheng, et al., 2025, Motie & Raahemi, 2024). Because standard GNNs are correlational, if a GNN learns that just being connected to a known fraudster is a strong signal of fraud, it might incorrectly flag a legitimate user who is a victim of fraud. Causal GNNs are a developing class of models that aim to integrate causal reasoning into the graph learning process and distinguish causal relationships from co-occurrence (Nguyen, et al., 2025). CaT-GNN (Causal Temporal Graph Neural Network) models were designed to address the problem of confounding in transaction graphs (Duan, et al., 2024, Sultana, et al., 2025). CausalFD (Song, Y. et al., 2024) uses causal inference to uncover the stable, underlying preferences or intentions of users, even as their superficial behaviours change. Causal GNNs hold the potential to detect sophisticated fraud rings as well as identify the influential actors within them, providing much more actionable intelligence for investigators (Nguyen, et al., 2025).

2.6 Drift in Machine Learning

The utility of a machine learning model lies in its ability to generalize from the data on which it was trained to new, unseen data encountered post deployment in a production environment. A key assumption is that the model will later be exposed to new data having the same distribution as the training data (Bickel, et al., 2009). This assumption makes it inherently vulnerable in real-world systems, which are rarely static. In fraud detection, this evolution is often due to fraudsters constantly devising new tactics to evade current mitigation techniques, rendering old pattern recognition systems obsolete (Bayram, et al., 2022). Covariate shift (or data drift) is a change in the distribution of the input features to the model which does not change the relationship between the features and target variable, leading to poor generalization (Bayram, et al., 2022). Concept drift occurs when the relationship between a model's input features and the predicted variable changes over time. Drift often significantly degrades the performance of machine learning models (Gama et al. 2014, Liu, et. al, 2018). Drift may occur in varying temporal patterns, further impacting model degradation. Sudden (or Abrupt) Drift occurs when a new concept or distribution replaces the old one almost instantaneously. Gradual (or Incremental) Drift is a slow and continuous transition from the old

concept to the new concept. Both concepts coexist for a period, with the new concept gradually becoming more prevalent. Recurring Drift occurs when previously seen concept reappears after a period of absence and is often tied to cyclical or seasonal phenomena (Deng, et al., 2025).

2.7 Drift detection techniques

Drift detection techniques are used to identify model drift and are broadly classified into statistical monitoring techniques, window-based methods, and learning-based or model-based approaches. Early work in drift detection primarily used statistical tests to monitor changes in model error rates. These include:

- Drift Detection Method (DDM) (Gama et al. 2004) uses the binomial distribution to track the classifier's error rate, DDM signals drift when the deviations are statistically significant.
- Early Drift Detection Method (EDDM) improves upon DDM by focusing on the distance between classification errors, making it more sensitive to gradual drifts (Baena-Garcıa et al.,2006).
- HDDM (Pearson and A-Test variants) applies statistical hypothesis testing, using Hoeffding's inequality or the A-test, to compare distributions between sliding windows (Frías-Blanco et al., 2015)
- ADWIN (Adaptive Windowing) (Bifet and Gavaldà, 2007) introduces an adaptive window size to detect changes in data distribution by maintaining two sub-windows and applying a Hoeffding bound to identify statistically significant differences. The method is effective in environments with both abrupt and gradual drift.
- Transaction Window Bagging (TWB) models have been used to handle concept drift (Somasundaram & Reddy, 2019). It is a parallelised, incremental bagging ensemble which handles imbalance by creating bags with all minority instances and sampled, temporally overlapping majority instances. It dynamically incorporates new, recent data bags with golden ratio-based temporal weights and prunes older models to handle concept drift (Dal Pozzolo et al., 2015, Topal et al., 2025).

2.7.1 Quantifying Data Distribution Shifts

- Kullback-Leibler (KL) Divergence is a non-symmetric measure of how one probability distribution is different to a second, reference probability distribution. A higher KL divergence indicates a greater dissimilarity between the two distributions. (Van Erven and Harremos (2014).
- Population Stability Index (PSI) is widely used in credit risk modelling and fraud detection to quantify shifts in a variable's distribution between two samples. PSI provides a single numer-

ical value that quantifies the magnitude of the shift, with established thresholds indicating no significant, moderate, or significant change (Yurdakul & Naranjo, 2020).

- Kolmogorov-Smirnov (KS) Statistic quantifies the maximum distance between the empirical cumulative distribution functions (CDFs) of two samples. It is used to test if two samples are drawn from the same distribution, or if a sample is drawn from a reference distribution. A larger KS statistic indicates a greater divergence between the distributions, making it valuable for detecting shifts in individual feature distributions (Porwik & Dadzie, 2022).

2.7.2 Quantifying Causal Structure Shifts

- Structural Hamming Distance (SHD) is a metric used to compare the structural differences between two causal graphs (DAGs). It counts the minimum number of edge operations (addition, deletion, or reversal of direction) required to transform one graph into another. A lower SHD indicates greater similarity between the two causal structures. While simple and intuitive, SHD treats all edge changes equally, even if some changes might have a larger impact on causal effects than others (de Jongh & Druzdzel, 2009).
- Structural Intervention Distance (SID) provides a more comprehensive measure of dissimilarity between two causal graphs. Instead of just counting edge operations, SID considers the differences in the sets of interventional distributions implied by each graph. It quantifies how many causal effects would differ if there was intervention on variables in the graph. A lower SID indicates that the two graphs imply very similar causal consequences, making it a more robust measure when the goal is accurate causal inference (Peters & Buhlmann, 2015).

3 RESEARCH METHODS

3.1 Synthetic Data Generation and Drift Simulation

Synthetic generation affords full control over the data-generating process, provides an unambiguous ground-truth causal structure, and enables repeatable simulation of prescribed drift scenarios without ethical or privacy concerns. To structure the data generation process and progressively increase complexity, the study will follow a two-phase synthetic data design.

3.1.1 Simple synthetic dataset

The dataset will be created using a small, hand-crafted DAG with a clearly defined causal structure and a limited number of variables. This will serve as a controlled environment for validating the performance of causal discovery algorithms under idealised, noise-controlled conditions.

3.1.2 Domain representative dataset

This dataset will be constructed based on domain knowledge of telecommunications fraud and will reflect real-world feature complexity, non-linearity, noise, and realistic scale. The known causal structure of this dataset will still be explicitly defined via a structural causal model, allowing for consistent and interpretable evaluation across both phases.

1. Structural Causal Model (SCM) Definition: The synthetic dataset will be generated from a domain-informed Structural Causal Model (SCM), represented as a Directed Acyclic Graph (DAG).

   $$X_i = f_i(\text{Parents}(X_i)) + \epsilon_i$$

   where:

   - f_i is a deterministic function capturing both linear and non-linear dependencies on its direct causes (parent nodes in the DAG),
   - ϵ_i is an independent noise term.

2. Feature distributions: To reflect realistic behavioural complexity in fraud-related activity, the functional forms that will be incorporated into the dataset are linear transformations for simple relationships, sigmoid (logistic) functions to simulate saturation effects, polynomial and interaction terms to model multiplicative or joint effects, logarithmic or exponential functions to reflect accelerating or diminishing returns, sinusoidal terms to encode periodic temporal effects and threshold-based activations to reflect triggering of fraud mechanisms only after certain limits.

3. Noise Injection: Noise will be added to emulate real-world variability. This will include Gaussian noise for continuous variables, Laplace noise to introduce occasional heavy-tailed outliers that mimic spiky or bursty transaction behaviours and categorical noise in the form of probabilistic label flipping for fraud outcomes, reflecting misclassification due to human or system error.
4. The sample dataset size will be selected to ensure sufficient statistical power to train and evaluate the models and reflect a realistic class imbalance for fraud use cases.

3.1.3 Representative Features of Synthetic Dataset

Each sample represents an account change attempt (SIM swap or new subscription) labelled as fraudulent or not fraudulent, and the features represent events prior to and after the event. This list of attributes to be included is not exhaustive and will be adapted as the research progresses.

- Interaction data: Prior interactions and requests with customer service channels.
- Transactions: Airtime or data bundle transfers and new orders.
- IMEI (device) and IMSI (SIM card) relationships: the number of SIM cards associated with the device used to request the SIM swap, and vice versa obtained through Call Detailed Records (CDRs).
- Location data: originating cell tower location and geographic spread of activity.
- Fraud indicators: prior activity by the user which was deemed fraudulent or suspicious through investigation, including repeated failed authentication attempts, locking of the line or blocking of the device on the network and historical fraud links.

3.2 Drift Simulation

Drift will then be simulated on the baseline datasets (simple and domain representative) to enable the evaluation of model robustness under distributional shifts. It will be introduced iteratively to produce a sequence of temporally ordered datasets, allowing for fine-grained analysis of performance degradation and causal stability over time. The simulations will be designed to isolate different types of drift in order to analyse their distinct impacts.

Feature Drift (Covariate Shift) simulation techniques

- Mean shift and variance shift will be simulated through sampling from different parametrisations of a distribution, adding constants to feature columns and multiplying features by a factor.

- To simulate a more fundamental change in the data, the underlying probability distribution of a feature will be changed entirely.
- Feature correlations will be altered by changing the covariance matrix to simulate domain changes where relationships between variables evolve.
- Class proportions will be modified in the test set by increasing or decreasing the instances of a specific class in the training data to simulate label shift.
- Sub-population sampling will be simulated by under-sampling or oversampling.
- Group exclusion or range removal: Entire categories from the data will be dropped or data from a certain range of a continuous variable will be removed. Label proportions will be balanced to prevent concept drift.

Concept Drift simulation techniques

- Boundary modification: The geometry of the decision boundary in the dataset will be modified.
- Feature importance drift: The coefficients that determine the importance of the features in the classification will be modified to simulate a shift in fraud strategies.
- Rule Substitution: The underlying causal function for the positive class will be replaced with a new rule at a specific time step to simulate the emergence of a novel fraud tactic.
- Structural drift: Modifications to the underlying DAG (edge addition, deletion, reversal) reflecting evolving fraud mechanisms.

Temporal Patterns of Drift

Four temporal patterns will be simulated across both feature drift and concept drift.

- Sudden drift: Abrupt changes will be introduced at a fixed time step. For feature drift, the distribution of selected variables will be changed. For concept drift, the underlying labelling function will be replaced with a new rule.
- Gradual Drift: A smooth transition in feature distributions will be simulated by blending old and new states over a defined time window. Concept drift will be simulated by progressively shifting the decision boundary and blending two labelling rules.
- Incremental drift: Stepwise modifications will be introduced at regular intervals. For feature drift, statistical properties (mean and variance) of key features will be adjusted incrementally. For concept drift, parameters such as feature weights or thresholds in the classification function will be adjusted in discrete steps.
- Recurrent drift: Cyclic patterns will be modelled by periodically alternating between previously observed feature distributions or labelling functions.

The output of this phase is a Python-based tool for generating synthetic causal datasets and simulating configurable data drift scenarios. The full procedure yields a sequence of datasets $\{D_0, D_1, \ldots, D_n\}$, accompanied by metadata specifying the exact drift parameters applied at each iteration.

3.3 Baseline Modelling

The baseline model performance for both correlational machine learning and causal models will then be established using the non-drifted dataset (D_0). Statistical models (Logistic Regression, Random Forests, Gradient Boosting and SVMs) will be trained and optimized through hyperparameter tuning (Grid Search or Randomized Search). The best-performing iteration will be selected based on predictive performance evaluated on a held-out test set using the appropriate metrics (confusion matrix, F1 score, recall, precision, AUC, and log loss). This baseline model will serve as the benchmark for evaluating subsequent performance changes due to drift. Given that the synthetic dataset is generated with a known ground truth causal graph, the baseline structural causal model will be established and its parameters optimized for causal inference and prediction. Causal discovery algorithms (PC, GES, FCI, and LiNGAM) will be applied to the synthetic dataset to evaluate their ability to recover the underlying causal structure without prior knowledge. The discovered graph will be compared against the known ground truth graph as a sanity check to validate that structural learning is accurate in the absence of drift. The output of this step will be the trained correlational model object and the causal model.

3.4 Iterative Drift Application and Evaluation

For each drifted dataset, the model performance degradation will be quantified relative to the baseline model metrics, while simultaneously measuring shifts in data distribution and causal structure. For the predictive machine learning model, the baseline model will be scored on each drifted dataset first without any retraining, and then with retraining. The predictive performance obtained at each iteration will then be compared to the baseline performance on the non-drifted dataset.

As with the predictive machine learning model, performance on each drifted dataset will be compared to the initial baseline. To evaluate the causal model, the discovered graphs for each drift iteration will be compared to the true causal graph using Structural Hamming Distance (SHD) and Structural Intervention Distance (SID), and the Average Treatment Effects (ATEs) will be compared to the true ATEs from the unchanged dataset. The extent of data drift will be quantified at each step using KL Divergence, Population Stability Index (PSI), and Kolmogorov-Smirnov (KS) to assess changes in feature distributions.

3.5 Tools and Libraries

Core Data Science and Numerical Libraries

- **NumPy**: For fundamental numerical operations, array manipulations, and the generation of synthetic data distributions.
- **Pandas**: For data manipulation, cleaning, and structuring.
- **SciPy**: For implementing statistical tests to verify data drift.
- **Custom Python Scripts**: Will be developed to generate synthetic datasets and introduce covariate and concept drift by altering data distributions and underlying data-generating functions.

Machine Learning Libraries

- **Scikit-learn**: For data preprocessing and training the baseline machine learning models
- **XGBoost and LightGBM**: For implementing state-of-the-art gradient boosting models.
- **TensorFlow and Keras/PyTorch**: For building and training custom neural network architectures.

Causal Discovery and Modelling

- **EconML**: For implementing advanced causal inference models, including meta-learners (S-Learner, T-Learner, X-Learner) and instrumental variable methods like DeepIV.
- **CausalML**: For implementing meta-learners and other causal methods.
- **DoWhy**: For specifying causal graphs, perform identification of causal effects, estimate these effects, and carry out robustness checks to validate causal conclusions
- **GRF (Generalized Random Forests)**: For implementing Causal Forests, designed for heterogeneous treatment effect estimation.

4 ETHICAL ISSUES

4.1 Synthetic Data Integrity

While the initial use of synthetic data mitigates privacy risks in the first phase of the study, the generated dataset must accurately reflect real-world fraud behaviour to ensure valid conclusions. The data generation process will be documented to ensure transparency. Domain expertise will be leveraged to validate the realism of the simulated dataset.

4.2 Data Privacy and Confidentiality

Depending on the results of the first phase of the study, real-world telecommunications data may be leveraged to further test the derived frameworks and methods. Any real world dataset used will be anonymised and any personal identifiable information (PII) will be removed to ensure compliance with POPIA. Data handling protocols will be established to ensure secure access storage, processing, and disposal.

4.3 Fairness and Bias Mitigation

The study will assess model performance across different sub-populations to prevent the introduction of bias. The use of causal models is expected to improve fairness by focusing on underlying mechanisms rather than spurious correlations.

4.4 Explainability and Responsible Use

While causal models offer improved interpretability compared to traditional black-box models, there may be a risk of misinterpretation or over-reliance on causal graphs. Robustness checks and sensitivity analyses will be performed to ensure that conclusions are valid and the findings will be communicated with appropriate caveats.

5 PROJECT SCHEDULE

The estimated project task progression is shown in Table 3. Iteration may be required depending on outcomes of key tasks such as model training which may impact planned time lines. Write up of the dissertation will occur concurrently with other tasks.

Table 3: Project Schedule

Task	**Start Date**
Phase 1: Data Generation & Drift Simulation	Sep 2025
1.1 Define SCM & Generate Baseline Data	Sep 2025
1.2 Simulate Drift Iterations D_i	Oct 2025
Phase 2: Model Training & Initial Evaluation	Oct 2025
2.1 Train Statistical Models	Oct 2025
2.2 Establish Causal Model Baseline	Nov 2025
Phase 3: Iterative Drift Application & Evaluation	Nov 2025
3.1 Score Static Models on D_i	Nov 2025
3.3 Quantify Drift & Graph Changes	Nov 2025
3.4 Causal Discovery on D_i	Dec 2025
Phase 4: Comparative Analysis & Interpretation	Mar 2026
4.1 Quantify Sustained Performance	Mar 2026
4.2 Analyse Degradation vs. Drift	Apr 2026
4.3 Assess Causal Stability & Accuracy	Apr 2026
Phase 5: Thesis Writing & Dissemination	May 2026
5.1 Draft Methodology & Results Chapters	May 2026
5.2 Discussion & Conclusion Drafting	Jun 2026
5.3 Review & Finalization	Jul 2026
5.4 Presentation Preparation	Aug 2026

6 REFERENCES

1. Adams, S., Griffiths, C., & Delahaye, J.-P. (2021). Root causes of process concept drift: A causality-based approach. International Journal of Business Process Integration and Management, 11(2), 145–158.

2. Babaei, K., Chen, Z. Y. & Maul, T., 2019. A Study of Fraud Types, Challenges and Detection Approaches in Telecommunication. Journal of Information Systems and Telecommunication, 7(4), pp. 248-261.

3. Baena-García, M., del Campo-Ávila, J., Fidalgo, R., Bifet, A., Gavaldà, R., and Morales-Bueno, R. (2006). Early Drift Detection Method. In Fourth International Workshop on Knowledge Discovery from Data Streams.

4. Bayram, F., Ahmed, B. S. & Kassler, A., 2022. From concept drift to model degradation: An overview on performance-aware drift detectors. Knowledge-Based Learning, Volume 245.

5. Bickel, S., Bruckner, M. & Scheffer, T., 2009. Discriminative Learning Under Covariate Shift. Journal of Machine Learning Research, Volume 10, pp. 2137-2155.

6. Bifet, A., & Gavaldà, R. (2007). Learning from Time-Changing Data with Adaptive Windowing. In Proceedings of the 2007 SIAM International Conference on Data Mining (pp. 443–448).

7. Carcillo, F, 2019. Combining Unsupervised and Supervised Learning in Credit Card Fraud Detection. Information Sciences.

8. Chen, Y., Zhao, C., Xu, Y. & Nie, C., 2025. Year-over-Year Developments in Financial Fraud Detection via Deep Learning: A Systematic Literature Review, s.l.: arXiv.

9. Cheng, D., Zou, Y., Xiang, S. & Jiang, C., 2025. Graph Neural Networks for Financial Fraud Detection: A Review. Frontiers of Computer Science, Volume 19.

10. Claassen, T., Mooij, J. M., & Heskes, T. (2013). Learning sparse causal models is not NP-hard. UAI 2013. https://arxiv.org/abs/1210.4862

11. Colombo, D., Maathuis, M. H., Kalisch, M., and Richardson, T. S. 2012. Learning high-dimensional directed acyclic graphs with latent and selection variables. The Annals of Statistics, 40(1), 294–321. https://doi.org/10.1214/11-AOS940

12. Dal Pozzolo, A. et al., 2015. Credit Card Fraud Detection and Concept-Drift Adaptation with Delayed Supervised Information. Killarney, Ireland, International Joint Conference on Neural Networks.

13. de Jongh, M. & Druzdzel, M. J., 2009. A Comparison of Structural Distance Measures for Causal Bayesian Network Models. Recent Advances in Intelligent Information Systems, pp. 443-458.

14. Deng, Z., Feng, Q. L. B. & Yen, G. G., 2025. Zilean: A modularized framework for large-scale temporal concept drift type classification. Information Sciences, 712(122134)

15. Duan, Y. et al., 2024. CaT-GNN: Enhancing Credit Card Fraud Detection via Causal Temporal Graph Neural Networks. arXiv preprint arXiv:2402.14708.

16. Ekeh, G. E. et al., 2022. Awareness of BVN, SIM swap and Clone Frauds: Methods and Controls. Science World Journal, 17(2), pp. 200-206.

17. Elwell, R., & Polikar, R. (2011). Incremental Learning of Concept Drift in Nonstationary Environments. IEEE Transactions on Neural Networks, 22(10), 1517–1531.

18. Estevez, P. A., Held, C. M. & Perez, C. A., 2006. Subscription fraud prevention in telecommunications using fuzzy rules and neural networks. Expert Systems Applications, 31(2), pp. 337-3334. , Fenstermacher, R., 2025. AI agent fraud: key attack vectors and how to defend against them. [Online] Available at: https://stytch.com/blog/ai-agent-fraud/ [Accessed 24 7 2025].

19. Frías-Blanco, I., del Campo-Ávila, J., Ramos-Jiménez, G., Morales-Bueno, R., Ortuno, F. M., & Krawczyk, B. (2015). Online and Nonparametric Drift Detection Methods Based on Hoeffding's Bounds. IEEE Transactions on Knowledge and Data Engineering, 27(3), 810–823.

20. Gama, J., Medas, P., Castillo, G., and Rodrigues, P. (2004). Learning with Drift Detection. In *Brazilian Symposium on Artificial Intelligence* (pp. 286–295).

21. Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. (2012). A Kernel Two-Sample Test. Journal of Machine Learning Research, 13(1), 723–773.

22. Hilas, C. S., 2009. Designing an expert system for fraud detection in private telecommunications networks. Expert Systems with Applications, Volume 36, pp. 11559-11569

23. Hu, M. et al., 2023. A Framework for Analyzing Fraud Risk Warning and Interference Effects by Fusing Multivariate Heterogeneous Data: A Bayesian Belief Network. Entropy, 25(6).

24. Hu, X. et al., 2022. BTG: A Bridge to Graph machine learning in telecommunications fraud. Future Generation Computer Systems, Issue 137, p. 274–287.

25. Jesus, S. et al., 2022. Turning the Tables: Biased, Imbalanced, Dynamic Tabular Datasets for ML Evaluation. s.l., 36th Conference on Neural Information Processing Systems.

26. Ke, G. et al., 2017. LightGBM: A highly efficient Gradient Boosting Decision Tree. Long Beach, CA, 31st Conference on Neural Information Processing Systems.

27. Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., and Zhang, G. (2018). Learning under concept drift: A review. IEEE Transactions on Knowledge and Data Engineering, 31(12), 2346–2363.

28. Misra, S., Thakur, S., Ghosh, M. Saha, S. K., 2020. An Autoencoder Based Model for Detecting Fraudulent Credit Card Transaction. Procedia Computer Science, Volume 167, pp. 254-262.

29. Moso, J. C. & Kenei, J. K., 2018. Credit Card Fraud Detection using Bayes Theorem. International Journal of Computer and Information Technology, 7(4), pp. 184-189.

30. Motie, S. & Raahemi, B., 2024. Financial fraud detection using graph neural networks: A systematic review. Expert Systems with Applications, Volume 240.

31. Mukhanov & Lev, 2008. Using Bayesian Belief Networks for credit card fraud detection. s.l., 26th IASTED International Conference on Artificial Intelligence and Applications.

32. Neal, B., 2020. Introduction to Causal Inference

33. Nguyen, L., Boersma, M. & Acar, E., 2025. Detecting Fraud in Financial Networks: A Semi-Supervised GNN Approach with Granger-Causal Explanations. Statistical Finance.

34. Nobel, S. et al., 2024. Unmasking Banking Fraud: Unleashing the Power of Machine Learning and Explainable AI (XAI) on Imbalanced Data. Information, 15(289).

35. Ogarrio, J. M., Spirtes, P., & Ramsey, J. (2016). A hybrid causal search algorithm for latent variable models. JMLR Workshop and Conference Proceedings.

36. Page, E. S. (1954). Continuous Inspection Schemes. *Biometrika*, 41(1/2), 100–115.

37. Pearl, J., 2000. Causality: Models, Reasoning, and Inference, Cambridge University Press.

38. Pearl, J. (2009). Causality: Models, Reasoning, and Inference (2nd ed.). Cambridge University Press.

39. Peters, J. & Buhlmann, P., 2015. Structural intervention distance (SID) for evaluating causal graphs. Neural Computation, 27(3), pp. 771-799.

40. Porwik, P. & Dadzie, B. M., 2022. Detection of data drift in a two-dimensional stream using the Kolmogorov-Smirnov test. Procedia Computer Science, Volume 207, pp. 168-175.

41. Psychoula, I. et al., 2021. Explainable Machine Learning for Fraud Detection. IEEE Computer Special Issue on Explainable AI and Machine Learning.

42. Pushkarenko, Y. & Zaslavskyi, V., 2024. Synthetic Data Generation for Fraud Detection Using Diffusion Models. Information and Security, 55(2), pp. 185-198.

43. Raghavan, P. El Gayar, N., 2019. Fraud Detection using Machine Learning and Deep Learning. Dubai, 2019 International Conference on Computational Intelligence and Knowledge Economy.

44. Reddy, A., 2024. Rise in telco fraud threatens digital trust in South Africa. [Online]

45. Shimizu, S., Hoyer, P. O., Hyvarinen, A. and Kerminen, A., 2006. A Linear Non-Gaussian Acyclic Model for Causal Discovery. Journal of Machine Learning Research, Volume 7, pp. 2003-2030.

46. Solomon Kullback and Richard A Leibler. On information and sufficiency. The annals of mathematical statistics, 22(1):79–86, 1951

47. Somasundaram, A. & Reddy, S., 2019. Parallel and incremental credit card fraud detection model to handle concept drift and data imbalance. Neural Computing and Applications, Volume 31, pp. 3-14.

48. Song, Y. et al., 2024. CausalFD: causal invariance based fraud detection against camouflaged preference. International Journal of Machine Learning and Cybernetics, Issue 15, p. 5053–5070.

49. Spirtes, P., Glymour, C. & Scheines, R., 2000. Causation, Prediction, and Search. 2nd ed. s.l.:The MIT Press.

50. Sultana, I., Maheen, S. M., Kshetri, N. & Zim, M. N. F., 2025. detectGNN: Harnessing Graph Neural Networks for Enhanced Fraud Detection in Credit Card Transactions. s.l., 2025 13th International Symposium on Digital Forensics and Security (ISDFS) (pp. 1-6). IEEE.

51. Topal, M., Bozanta, A., Erer, E. & Basar, A., 2025. Handling Concept Drift in Fraud Detection: A Replication Study. Calgary, The 38th Canadian Conference on Artificial Intelligence.

52. Transunion, 2024. Global Telecoms Hit Hard: 38.95 Billion Lost to Fraud in 2023. [Online] Available at: https://www.transunionafrica.com/blog/global-telecoms-hit-hard-by-fraud

53. Transunion, 2024. State of Omnichannel Fraud Report, Transunion. Available at: https://techcentral.co.z fraud-digital-trust-in-south-africa/255568/

54. Van Erven, T. and Harremos, P. (2014). Rényi divergence and kullback-leibler divergence. IEEE Transactions on Information Theory, 60(7):37973820.

55. Wang, C., Nie, C. & Liu, Y., 2025. Evaluating Supervised Learning Models for Fraud Detection: A Comparative Study of Classical and Deep Architectures on Imbalanced Transaction Data, s.l.: arXiv.

56. Webb, G. I., Hyde, R., Cao, H. & Nguyen, H. L., 2025. Characterizing Concept Drift. Data Mining and Knowledge Discovery.

57. Wijaya, M. G., Pinaringgi, M. F. & Zakiyyah, A. Y., 2024. Comparative Analysis of Machine Learning Algorithms and Data Balancing Techniques For Credit Card Fraud Detection. Ngaglik, Procedia Computer Science.

58. Yang, L., Cheng, J., Luo, Y., Zhou, T., and Zhang, X. (2025). Detecting and rationalizing concept drift: A feature-level approach for understanding cause–effect relationships in dynamic environments. Expert Systems with Applications, 260, 125365. https://doi.org/10.1016/j.eswa.2024.125365

59. Yurdakul, B. & Naranjo, J. D., 2020. Statistical Properties of the Population Stability Index. The Journal of Risk Model Validation.

60. Zhang, D., Bhandari, B. & Black, D., 2020. Credit Card Fraud Detection Using Weighted Support Vector Machine. Applied Mathematics, 11(12), pp. 1275-1291.

61. Zheng, X., Aragam, B., Ravikumar, P. K. Xing, E. P., 2018. Dags with no tears: Continuous optimization for structure learning. Advances in neural information processing systems, Volume 31.

62. Zhou, C. Paffenroth, R., 2017. Anomaly Detection with Robust Deep Autoencoders. Halifax, 23rd ACM SIGKDD International Conference.

7 Appendices

A GLOSSARY

- MSISDN (Mobile Station International Subscriber Directory Number) - the mobile subscriber's phone number including the international country code.
- IMEI (International Mobile Equipment Identity) - unique identifier for a GSM device
- IMSI (International Mobile Subscriber Identity) - unique identifier for a SIM card
- CLI (Caller Line Identifier) - the MSISDN of the caller
- Atoll ID - Identifier for cell tower
- Airtime Transfer - Method by which one subscriber on a telecommunications network can transfer airtime to another subscriber
- Airtime Loan - Service offered by telecommunications companies allowing a customer to buy airtime on credit for a service fee
- OTP (One Time Pin)
- Phishing: A type of cyber attack that uses deceptive emails or websites to trick individuals into revealing sensitive information, such as usernames, passwords, and credit card details
- Vishing: A type of cyber attack that uses phone calls or voice messages to deceive individuals into divulging personal and financial information.
- Smishing: A type of cyber attack that uses text messages (SMS) to trick individuals into revealing their sensitive information or downloading malware
- prepaid: A payment method where customers pay for their telecommunications services before using them. Customers are required to purchase airtime usage is deducted from that balance.
- postpaid: A payment method where customers are billed after they have used the telecommunications services.