MAPÚA MALAYAN
DIGITAL COLLEGE

A college under Mapúa Malayan Colleges Laguna

Intellectual Property Notice

This template is an exclusive property of **Mapua-Malayan Digital College** and is protected under **Republic Act No. 8293**, also known as the *Intellectual Property Code of the Philippines* (IP Code). It is provided solely for educational purposes within this course. Students may use this template to complete their tasks but may not **modify, distribute, sell, upload,** or **claim ownership** of the template itself. Such actions constitute copyright infringement under **Sections 172, 177, and 216** of the IP Code and may result in legal consequences. Unauthorized use beyond this course may result in legal or academic consequences.

Additionally, students must comply with the **Mapua-Malayan Digital College Student Handbook**, particularly with the following provisions:

- **Offenses Related to MMDC IT**:
  - **Section 6.2** – Unauthorized copying of files
  - **Section 6.8** – Extraction of protected, copyrighted, and/or confidential information by electronic means using MMDC IT infrastructure
- **Offenses Related to MMDC Admin, IT, and Operations**:
  - **Section 4.5** – Unauthorized collection or extraction of money, checks, or other instruments of monetary equivalent in connection with matters pertaining to MMDC

Violations of these policies may result in **disciplinary actions ranging from suspension to dismissal**, in accordance with the Student Handbook.

For permissions or inquiries, please contact MMDC-ISD at isd@mmdc.mcl.edu.ph.

Detecting Fraud in Insurance Financial Transactions Using Statistical and Machine Learning-Based Analytics

A Capstone Project Proposal

Presented to

Baluyot, Glenn
Tamayo, Aldrin John
Sta Cruz III, Armando

In Partial Fulfillment of the Requirements
for the MO-IT200D1 Capstone 1 Course

Presented by

Años, Shanne
Atienza, Trisha Mei
Ongleo, Christen Amiel O.
H3103

Bachelor of Science in Information Technology
Major in BSIT - Data Analytics

May 2025

ACKNOWLEDGEMENT

Type here your acknowledgements. This is where you and your group members express gratitude to individuals, organizations, or groups who contributed to the completion of your study.

Make sure that you indent and add one (1) space when you need to move on to another paragraph. Use appropriate language, proper conventions, and academic writing style. Definitely, delete these statements in red when you're going to type in your paper; follow proper format.

Read on to know some general reminders when writing your paper. Feel free to use bullets, charts, tables, or images when necessary. When stating a number, be consistent if you will use the word format (e.g., one, five, sixty-two) or the arabic numerals (e.g., 6, ½, 80%). Be mindful of your in-text citations, which should also be listed on the References page of your paper.

TABLE OF CONTENTS

CHAPTER I
INTRODUCTION

A. Background of the Study

The insurance industry serves as a vital pillar in safeguarding individuals, businesses, and economies against financial uncertainties. However, in recent years, it has increasingly become a target of fraudulent practices that compromise the integrity of its systems. These fraudulent activities—whether through inflated, manipulated, or entirely falsified claims—pose serious threats to insurers by causing financial losses and undermining policyholder confidence.

Globally, insurance fraud has escalated to alarming levels, resulting in billions of dollars in losses each year. As noted by Syamkumar et al. (2024), this growing trend forces insurers to raise premiums, ultimately impacting honest policyholders. Fraud affects all lines of insurance—such as health, property, motor vehicle, and life insurance—but is especially damaging in regions with emerging digital infrastructure, where detection mechanisms are still evolving.

In the Philippines, this issue is particularly critical. According to the Insurance Commission, there has been a noticeable increase in fraudulent claims, especially in the motor vehicle and healthcare sectors. Industry reports estimate that in 2022, local insurers incurred over PHP 950 million in losses due to fraud. Considering the Philippine insurance market's size—estimated at over PHP 300 billion—the economic implications of unchecked fraud are substantial. These challenges are further compounded by regulatory constraints, resource limitations, and the accelerating shift toward digital claim processing.

Historically, insurance companies have relied on manual audits and rule-based systems to identify suspicious claims. While these methods have been foundational, they are increasingly inadequate in addressing sophisticated and adaptive fraud tactics. Studies by Roy and George (2017) and Debener et al. (2023) have shown that traditional approaches often lead to operational inefficiencies, delayed responses, and higher rates of false negatives.

To address these limitations, modern fraud detection is shifting toward data-driven approaches, particularly those involving statistical and machine learning techniques. Machine learning (ML) algorithms offer dynamic, pattern-based analysis capable of identifying complex and previously unseen fraud behaviors. Bauder and Khoshgoftaar (2020) highlight how ML models continuously learn from data, allowing them to adapt to emerging fraudulent trends over time.

This study proposes the development of a **hybrid fraud detection framework** specifically designed for the Philippine insurance sector. The framework integrates logistic regression—a statistical method offering interpretability and robustness—with advanced machine learning models such as random forest and k-nearest neighbors, which provide higher predictive accuracy and adaptability. This combination aims to enhance both the transparency and effectiveness of fraud detection systems.

To evaluate this hybrid model, the study will utilize a mix of **real-world insurance data and synthetically generated claims to simulate fraud scenarios.** Model performance will be assessed using metrics such as accuracy, precision, and recall to ensure a comprehensive evaluation. By doing so, the research seeks to contribute a practical and scalable solution to the growing issue of insurance fraud in the Philippines, reinforcing trust and fairness within the local insurance industry.

B. Statement of the Problem

Despite advancements in digital claim processing, insurance fraud remains a persistent and costly challenge, particularly in developing markets like the Philippines. Fraudulent claims—ranging from staged accidents and inflated medical bills to the use of fake identities and falsified documentation—continue to bypass existing detection systems. These methods are often sophisticated and designed to exploit the limitations of static, rule-based algorithms and manual verification processes, which struggle to detect non-obvious or novel fraud patterns.

Traditional fraud detection systems are typically rigid and reactive. They rely on predefined rules and historical fraud indicators, making them ill-equipped to detect evolving fraud tactics that do not match known red flags. For example, fraudulent claimants may use slightly altered personal information or exploit loopholes in health and motor insurance policies to avoid detection. This inability to adapt contributes to high false-positive rates—where legitimate claims are flagged unnecessarily—resulting in operational inefficiencies, wasted investigative resources, and customer dissatisfaction. Conversely, false negatives allow fraudulent claims to go undetected, leading to financial leakage.

In the Philippine insurance sector alone, fraud-related losses are estimated to exceed PHP 950 million annually. These losses not only affect the profitability of insurers but also increase premium rates for consumers and erode public trust in the industry. With the local insurance market valued at over PHP 300 billion, even a small percentage lost to fraud represents a significant threat to financial sustainability and market credibility.

Existing fraud detection solutions in the local context often lack adaptability, accuracy, and scalability. They are not designed to learn from emerging fraud trends, nor can they handle large-scale datasets in real time. This creates a critical gap in the industry's ability to proactively detect and respond to fraudulent activities.

This study addresses the pressing need for a more dynamic and intelligent fraud detection approach by proposing a hybrid framework that integrates statistical analysis with machine learning techniques. While the proposed methodology is detailed elsewhere in the paper, the core problem this research aims to address is the **inadequacy of current fraud detection systems to effectively identify complex and evolving fraudulent patterns** in the Philippine insurance industry. Tackling this issue is essential to reducing economic losses, improving operational efficiency, and restoring stakeholder confidence.

C. Research Objectives

This study aims to **design, develop, and evaluate** a hybrid fraud detection model that integrates traditional statistical methods—specifically z-score, Benford's law, time-based features, and rule-based flags—with machine learning algorithms such as logistic regression and random forest, in order to improve the detection accuracy and operational efficiency of identifying fraudulent insurance claims.

Specifically, the study seeks to:

1. **Assess current fraud detection practices** among Philippine insurance providers through rating-scale surveys and semi-structured interviews with claims analysts, IT personnel, and operations managers, to identify key pain points, system gaps, and opportunities for improvement.
2. **Acquire, clean, and preprocess** a structured dataset of labeled insurance claims by applying transformation techniques (e.g., normalization, encoding), feature engineering (e.g., fraud flags, time features), and data balancing (e.g., SMOTE) to ensure high-quality model inputs.
3. **Design, develop, and compare** multiple fraud detection models—including logistic regression, random forest, and k-nearest neighbors—by training them on preprocessed data and identifying behavior patterns linked to fraud.
4. **Evaluate model performance** using measurable classification metrics such as:
    a. **Accuracy (% of correct classifications)**
    b. **Precision (minimizing false positives)**
    c. **Recall (maximizing fraud detection)**
    d. **F1-score (balance between precision and recall)**
    e. **AUC-ROC (overall discrimination ability)**

    Models will be selected based on their performance and generalisability to unseen data.

D. Research Questions

To guide the study, the following research questions are posed:

1. Which statistical and machine learning techniques demonstrate the highest effectiveness in detecting fraudulent insurance claims, as measured by classification metrics such as accuracy, precision, recall, F1-score, and AUC-ROC?
2. What fraud indicators or patterns can be learned from claims data?
3. How does the proposed hybrid model conceptually compare to traditional detection approaches—such as manual audits and rule-based systems in terms of expected accuracy, reduction of false positives, and operational efficiency?
4. Which data preprocessing and feature selection techniques have the greatest impact on model accuracy and generalisability?
5. How can the proposed model be integrated into real-world insurance workflows to support faster and more accurate fraud detection?

E. Significance of the Study

In an industry where billions are lost annually to fraudulent activity, detecting deception in insurance claims is no longer a technical luxury — it's a business necessity.

This study directly addresses that urgency by proposing a fraud detection model grounded in both statistical and machine learning analytics, offering a practical and scalable solution to a decades-old challenge.

For Insurance Providers

This study introduces a deployable hybrid fraud detection framework designed to significantly reduce undetected fraud — a leading cause of multi-million peso losses in the Philippine insurance sector. Beyond financial savings, the model has the potential to transform operations by shifting from reactive, manual reviews to proactive, automated fraud detection.

By adopting this system, insurers can:

- **Minimize fraudulent payouts** and protect their bottom line
- **Unlock early warning systems** through advanced data pattern analysis
- **Reduce manpower load** by automating initial fraud checks
- **Enhance compliance** with transparent and auditable decision logic
- **Gain a competitive edge** by offering faster and more accurate claims processing

In high-volume environments, even a modest increase in detection precision can translate into millions saved annually and more strategic resource allocation.

For Customers and Policyholders

Fraud doesn't just hurt insurers — it penalises honest customers through higher premiums and delayed claims. This study contributes to:

- **Faster and fairer claim processing** by reducing false positives
- **Improved customer experience**, with less friction and suspicion
- **Greater trust** between providers and policyholders, especially in critical moments like health emergencies or natural disasters

By improving fraud detection precision, legitimate claimants are served better, faster, and more respectfully — fostering long-term customer loyalty and satisfaction.

For the Philippine Economy

A stronger, fraud-resilient insurance industry contributes to broader **economic stability** and **consumer confidence**. When fraud is curbed, insurance firms can offer **lower premiums**, invest more in service innovation, and extend coverage to underinsured sectors. In the long term, this contributes to a more inclusive financial system and supports national resilience during crises.

For the Data Science and Analytics Field

This project demonstrates the real-world impact of machine learning in a high-stakes, regulated domain. Its key technical contributions include:

- **Integration of SMOTE** (Synthetic Minority Oversampling Technique) to address class imbalance
- **Use of SHAP** (SHapley Additive exPlanations) for model interpretability and feature attribution
- **Application of the CRISP-DM** (Cross-Industry Standard Process for Data Mining) framework to structure and scale development
- And **comparative analysis** across traditional statistical models and ensemble ML algorithms to identify the best-performing approach.

Innovation emerges from the synergy of these methods. SMOTE ensures the models don't overlook minority (fraud) cases, while SHAP makes the decisions of complex ML models explainable to human reviewers — a vital requirement in regulated industries. CRISP-DM allows this entire workflow to be aligned with business goals and iteratively improved. Together, these components create a model that is not only **accurate**, but also **transparent**, **scalable**, and **ready for operational deployment**.

This approach showcases how thoughtful integration of techniques can bridge the gap between data science experimentation and industry-grade, ethically sound solutions — offering a blueprint for analytics practitioners tackling fraud and other high-risk use cases.

In a world where fraud tactics evolve faster than legacy systems can respond, this project offers a data-driven, intelligent response. One that directly reduces financial risk, enhances customer trust, and future-proofs fraud prevention systems.

This is not just a technical model — it's a strategic leap forward that the insurance industry can no longer afford to ignore.

F. Scope and Limitations

This study focuses on the design, development, and offline evaluation of a hybrid fraud detection framework that integrates both statistical methods (e.g., logistic regression, correlation analysis) and machine learning classification algorithms (e.g., decision trees, support vector machines, ensemble models). The scope includes detecting fraudulent activities within key insurance financial transactions such as claims processing, premium refund requests, and policyholder benefit disbursements. And also includes the collection, preprocessing, and analysis of **real-world insurance claim datasets comprising transactional data (e.g., claim amounts, dates, claim types), policyholder information (e.g., anonymised demographic details, policy type), and claim narratives or descriptions sourced from a local insurance company in the Philippines**, subject to availability and anonymisation protocols.

The tool we will be using for this research is JASP, and the framework will be trained and tested on retrospective, labeled data — containing both fraudulent and non-fraudulent claims. It will be evaluated using industry-standard classification metrics such as **accuracy, precision, recall, F1-score**, and **AUC-ROC**. Additionally, qualitative insights will be gathered through pre-assessment surveys and interviews with insurance stakeholders to contextualize current fraud detection challenges and inform the model design.

The study focuses solely on **offline performance assessment**. It does not include real-time system deployment or full-scale integration into live production systems.

Furthermore, while the study assumes access to anonymised datasets, it does not explicitly cover ethical, legal, or regulatory issues surrounding data use, such as **data privacy laws (e.g., Philippine Data Privacy Act)** or algorithmic fairness. These are acknowledged as vital areas for future research and implementation planning.

Key limitations of this study include:

- **Potential constraints in accessing complete or representative datasets** from the partner insurance company, which may impact the generalisability of the findings;
- **Imbalanced class distribution and data noise**, which are typical of fraud datasets and may influence model training and evaluation;
- The **absence of real-time validation**, which limits the assessment of the model's responsiveness and adaptability in operational settings.
- Since the data might be **specific or synthetic,** the finding's generalizability to all insurance companies or all types of fraud might need further validation.

Despite these limitations, the study aims to demonstrate the viability and effectiveness of a hybrid analytical approach in improving fraud detection in insurance claims, with the goal of informing future adoption in Philippine insurance workflows.

CHAPTER II
REVIEW OF RELATED LITERATURE

In the dynamic domain of financial and insurance systems, fraud detection remains a pressing concern due to the increasing volume and complexity of fraudulent activities. Traditional methods such as manual audits and rule-based systems, although widely adopted, have demonstrated numerous shortcomings in adapting to the sophisticated and evolving nature of fraud schemes. In response, the field has seen a surge in data-driven approaches—particularly those leveraging statistical models and machine learning (ML)—which offer both predictive power and operational scalability. This chapter provides a comprehensive review of current literature surrounding the limitations of conventional systems, the capabilities and interpretability of advanced ML models, trade-offs in performance metrics, data preparation strategies, and development frameworks. These thematic areas frame the capstone's research direction and directly inform its methodological choices, especially the development of a hybrid fraud detection system grounded in both statistical transparency and machine learning efficiency.

I. Traditional Statistical Methods Used for Fraud Detection

Insurance fraud continues to pose persistent challenges for providers, prompting the exploration of advanced analytics while maintaining a strong reliance on traditional statistical methods. These conventional approaches—known for their interpretability, regulatory compliance, and cost-effectiveness—remain relevant in operational settings, particularly where explainability and expert validation are non-negotiable. Unlike black-box ML models, statistical techniques allow domain experts to trace decisions, making them suitable for early-stage filters and transparent reporting. This section synthesizes four key statistical techniques widely applied in fraud detection: **Z-score analysis**, **Benford's Law deviation scores**, **manual rule flags**, and **time-based behavioral features**.

A. Z-Score Analysis

Z-score analysis is a statistical technique used to identify outliers by measuring how far a data point deviates from the mean, in terms of standard deviations. It is particularly useful in fraud detection, where fraudulent transactions often fall far outside normal behavioral patterns.

Bhavani and Amponsah (2017) compared the Beneish M-score and Altman Z-score models for detecting financial reporting fraud. While the Beneish model focused on earnings manipulation, the Altman Z-score—originally for bankruptcy prediction—proved more effective in flagging unusual financial behaviors. Its sensitivity to working capital changes, retained earnings, and market value made it adaptable beyond its original purpose, especially in detecting anomalies that may suggest fraud.

Labbaf (2023) applied the Z-score method to a large Kaggle credit card fraud dataset (284,807 transactions, with only 492 labeled as fraud). By flagging outliers during preprocessing, the study used Z-scores to enrich input for a downstream ML classifier. This led to strong performance improvements: precision (0.91), recall (0.82), and F1-score (0.86). Notably, this was achieved without using oversampling techniques, showing Z-score's potential to handle class imbalance efficiently while preserving data integrity.

Together, these studies illustrate how Z-score analysis offers both simplicity and interpretability—qualities essential in regulated fields like insurance. It serves as a fast, low-resource method to flag suspicious entries for further review or downstream modeling. However, it also has limitations. Z-scores assume normal distribution, which may not reflect real-world transaction data. They can also miss more nuanced fraud patterns that don't show up as statistical outliers.

In this capstone study, Z-scores will be calculated on numeric claim features to identify outliers as part of a hybrid fraud detection approach. These scores will be used as additional features in classification models (e.g., Logistic Regression, Random Forest), combining interpretability with machine learning power. This approach addresses a gap in existing research—where Z-score is often used for outlier filtering but rarely examined for its additive value as a model input in a fraud analytics pipeline.

B. Benford’s Law Deviation Score

Benford’s Law describes the expected frequency distribution of first digits in naturally occurring datasets, where lower digits (especially 1) appear more frequently than higher ones. It is widely used in forensic accounting and auditing because fraudulent or manipulated data often deviates from this pattern, making it a useful tool in financial anomaly detection.

Yap and Lai (2024) applied Benford's Law to Malaysian healthcare insurance claims to test whether claim categories conformed to expected digit distributions. Using chi-square, MAD, and Z-tests, they found that while hospitalization fees aligned with Benford’s Law, outpatient claims showed significant deviation—suggesting potential manipulation. The study highlighted Benford’s Law as a cost-effective early filter for fraud audits. However, it also noted that some deviations might result from normal pricing structures, not necessarily fraud, making human validation still essential.

Covacci (2025) explored a more integrated approach by embedding Benford deviation scores into machine learning classifiers—Logistic Regression, Random Forest, and K-Nearest Neighbors. Using the European credit card fraud dataset, Covacci generated features based on digit deviations and found that including them improved precision and AUC by 4–6%. This showed how statistical explainability can enhance model performance in hybrid systems.

Both studies reinforce the value of Benford’s Law in fraud detection—whether as a standalone red flag or a feature engineered into ML models. Its mathematical transparency and low computational cost make it ideal in regulated, resource-limited settings. However, its effectiveness depends on having datasets with wide, naturally distributed values. In constrained datasets (like fixed premiums or capped benefits), Benford’s assumptions may not apply, leading to false positives.

For this capstone, Benford deviation metrics will be used as part of the feature engineering process—alongside Z-scores and time-based indicators—to improve fraud detection in insurance claims. This addresses the gap where Benford’s Law is often applied in isolation. Integrating it into a hybrid model supports the study’s goal of building a system that balances accuracy, interpretability, and real-world applicability in the Philippine insurance sector.

C. Manual Rule Flags

Manual rule-based systems rely on predefined logic—such as "if claim amount exceeds X, flag as suspicious"—to identify potentially fraudulent activities. These systems are grounded in domain expertise and are commonly used as the first line of defense in fraud detection, especially in organizations where human oversight and regulatory compliance are critical.

The International Association of Insurance Supervisors (IAIS, 2022) outlined a global set of practical fraud indicators frequently used in insurance supervision. These include inconsistencies in submitted documents, unusually large claim amounts, rapid follow-up claims, and connections to previously flagged individuals or institutions. While the IAIS framework is not based on a specific dataset, it reflects international best practices and emphasizes the importance of manual rule flags in audit-heavy or low-tech environments—such as government-run programs or emerging insurance markets.

Islam, Haque, and Rezaul Karim (2024) developed a rule-based fraud detection model using transaction logs from a private fintech company. The model incorporated logic rules targeting behavioral anomalies like sudden spending spikes, high transaction frequency, and mismatches with a user's historical activity. Even without resampling techniques or advanced classifiers, the model achieved a high specificity rate (94%) and a balanced accuracy score of 87%. This demonstrates the continued effectiveness of manual rules when built on real-world behavioral knowledge and historical data patterns.

However, manual rule systems come with significant limitations. They are static by nature—requiring constant updates to remain relevant—and often generate high false positives. As fraud tactics evolve, hard-coded rules may become outdated or insufficient to detect subtle, complex schemes. Moreover, excessive reliance on rigid rules can slow down claims processing and reduce customer satisfaction due to unnecessary red flags.

In this capstone project, manual rule logic will not be used as a standalone detection system but will instead inform the creation of **interpretable features**. For example, domain-informed thresholds such as “claim amount > ₱500,000 and filed after 11PM” will be converted into binary flags like `high_risk_flag = 1`. These rule-based indicators will then be incorporated into the machine learning pipeline, combining expert knowledge with the adaptability of statistical models. This strategy ensures the benefits of manual rules—low cost and high interpretability—while addressing their limitations through data-driven learning. It supports the capstone’s aim to develop a hybrid fraud detection framework that is scalable, transparent, and tailored for real-world insurance workflows in the Philippines.

D. Time-Based Features

Time-based features capture behavioral patterns using timestamps—such as the time of day, day of the week, or gaps between transactions. These features are especially valuable in fraud detection because fraudulent activity often happens during off-peak hours, in bursts, or follows unusual timing patterns that differ from normal user behavior.

Li et al. (2019) introduced a Time Attention-Based Fraud Transaction Detection Framework (TAFDT), which combined users' static attributes (e.g., age, gender, transaction type) with time-based behavior such as transaction hour and time intervals between transactions. The model leveraged a temporal attention mechanism within a neural network architecture to highlight important sequence patterns associated with fraud. Tested on the Ant Financial dataset, the model achieved a high AUC of 0.941—showing that temporal cues can significantly boost detection performance, especially in sequential data settings.

Vivek et al. (2023) focused on real-time fraud detection in ATM transactions using Apache Spark Streaming paired with a Random Forest classifier. Key time-derived features included transaction hour, number of withdrawals in the past 24 hours, and periods of inactivity. The model achieved 93% accuracy and was capable of flagging suspicious activity within seconds. Their implementation demonstrated how simple timestamp features, when processed in real-time environments, can improve responsiveness and fraud detection outcomes—particularly in high-volume financial systems.

Both studies underscore the strength of time-based features in revealing subtle behavioral changes that static variables may miss. Unlike rule-based flags that require manual tuning, time features offer dynamic insights into transaction sequences and user behavior patterns. However, their effectiveness often depends on advanced models that can capture temporal dependencies—such as attention mechanisms or sequence-based classifiers. When used with simpler models, these features may not reach their full potential.

In this capstone study, time-based features will be engineered from timestamp data in insurance claims. This includes extracting `claim_hour`, `day_of_week`, `time_since_last_claim`, and `claims_in_last_30_days`. These variables will be added to the feature set used by machine learning classifiers (e.g., Logistic Regression, Random Forest). The goal is to detect behavioral anomalies such as frequent claims within short time windows or submissions during odd hours. Integrating these indicators strengthens the system's ability to detect fraud patterns while maintaining interpretability, aligning with the capstone's goal of building a practical, hybrid detection system for the Philippine insurance industry.

The following synthesis summarises the strengths, limitations, and implementation of the reviewed techniques, and explains how they directly support the capstone's methodology.

Synthesis

Traditional statistical methods remain valuable in fraud detection for their simplicity, transparency, and alignment with regulatory needs. Z-scores help flag unusual claim amounts; Benford's Law detects digit-level anomalies; manual rules use expert logic; and time-based features highlight suspicious patterns in timing or frequency.

However, each method has its limits. Z-scores assume normality, Benford's Law needs naturally distributed data, manual rules lack adaptability, and time-based insights require models that can process sequences. More importantly, few studies explore their combined use within machine learning pipelines.

This capstone addresses that gap by transforming these traditional methods into engineered features—feeding them into models like Logistic Regression and Random Forest. The goal: a hybrid system that is accurate, interpretable, and practical for real-world use in Philippine insurance fraud detection.

Table 1. Summary of Traditional Statistical Methods for Fraud Detection

Method	Key Strengths	Use Case	Limitations
Z-Score Analysis	Simple, interpretable; good for outlier detection	Flags anomalous transactions that deviate from statistical norms	Assumes normal distribution; may miss subtle fraud patterns
Benford's Law	Mathematically grounded; useful for fabricated data detection	Flags digit anomalies in claim amounts	Sensitive to dataset size; not ideal for capped or fixed-value datasets
Manual Rule Flags	Cost-effective; based on expert logic; aligned with compliance	Flags transactions using hard-coded thresholds and known red flags	Static, hard to scale; prone to high false positives; requires manual tuning
Time-Based Features	Captures behavioral shifts; strong in real-time or sequential setups	Detects suspicious timing (e.g., bursts, odd hours, rapid claims)	Requires time-aware models; less useful in static or rule-only systems

II. Limitations of Traditional Fraud Detection Systems

Despite decades of use, traditional fraud detection systems in the insurance industry—primarily rule-based engines and manual audits—are increasingly misaligned with the scale, speed, and complexity of modern fraud schemes. The growing digitisation of insurance processes has outpaced these legacy systems, exposing critical gaps in adaptability, accuracy, and scalability. This section synthesises the limitations identified in recent literature and draws clear connections to the motivations behind the capstone's proposed hybrid AI-based detection framework.

One consistent theme across studies is the rigidity of static rule-based systems. These systems operate on predefined thresholds and if-then rules, often crafted based on historical fraud cases. While initially effective, their performance deteriorates over time as fraudsters develop tactics that avoid known triggers. Aparício et al. (2020) and Kamalapurkar and Sharma (2025) both highlight how the inability of static systems to evolve with fraud patterns results in declining detection accuracy and growing rule maintenance costs. These systems require continuous manual updates to remain relevant, a process that is both time-consuming and error-prone. This inflexibility supports the capstone's choice to integrate machine learning models, which learn dynamically from data and adapt to new fraud behaviours without constant manual recalibration.

Equally concerning is the issue of high false positive rates. Studies by Rohn (2022) and Infosys Limited (2024) point out that conventional systems tend to flag a disproportionately large number of legitimate claims as suspicious. This creates significant inefficiencies by overwhelming fraud investigation units and delaying claims for honest customers—damaging both operational throughput and user experience. From a resource allocation perspective, this misdirection of investigative effort dilutes the focus on actual fraud cases. To address this, the capstone emphasises the use of precision-oriented models, such as logistic regression and random forest, which can be optimised to strike a balance between false positives and fraud detection sensitivity using threshold tuning and resampling techniques like SMOTE.

Another pressing limitation is the lack of scalability in manual audits and rule-based engines. As transaction volumes surge in the insurance sector—especially with the rise of digital submissions—human analysts cannot feasibly process each case in real-time. Infosys (2024) underscores how such systems become bottlenecks, particularly in emerging markets with rising claim volumes. Without automation, the risk of oversight increases, allowing sophisticated fraud schemes to pass undetected. This reality directly informs the capstone's objective to leverage automated, batch-based machine learning workflows that can process large datasets quickly while maintaining consistency in judgment.

Additionally, delayed fraud detection remains a structural flaw in traditional systems. Manual audits typically operate in a post-facto manner, identifying fraud only after claims are paid and losses incurred. Kamalapurkar and Sharma (2025) warn that this reactive model limits insurers' ability to intervene early, allowing fraudulent payouts to accumulate before corrective measures are taken. In contrast, the proposed capstone solution seeks to build proactive detection capability by training models on time-based features and patterns indicative of fraud, such as anomalous claim timing, suspicious payout velocity, or irregular policy behaviour. These engineered features not only enhance model accuracy but also reduce the lag between suspicious activity and action.

Collectively, these limitations provide a compelling justification for shifting from legacy systems to data-driven fraud detection models that are adaptive, scalable, and interpretable. While rule-based approaches laid the foundation for fraud analytics, current challenges demand smarter solutions capable of evolving alongside fraud tactics. The proposed capstone responds to these gaps by designing a hybrid framework that combines the interpretability of logistic regression with the adaptive learning power of random forest classifiers—optimised using real insurance data and validated through industry-standard performance metrics such as precision, recall, F1-score, and AUC-ROC.

In sum, literature shows a clear consensus: traditional methods alone are insufficient in today's digital insurance environment. Their shortcomings—static logic, high false positives, scalability constraints, and lagging detection—underscore the urgent need for hybrid fraud analytics systems that pair automation with statistical rigour. This capstone builds precisely on that need, contributing a context-sensitive, high-performance model designed to support Philippine insurers in curbing fraud and regaining operational control.

III. The Role of Explainable AI in Enhancing Trust

As machine learning models grow in complexity and adoption, the need for transparency and interpretability—particularly in high-stakes domains like insurance fraud detection—has become increasingly critical. Unlike traditional statistical methods, many advanced ML models, especially neural networks and ensemble methods, often function as "black boxes," making it difficult for users to understand how specific predictions are made. This opacity poses challenges not only in regulatory compliance but also in building trust among domain experts, such as insurance analysts, who are required to justify or act on algorithmic decisions.

To address this, recent literature has emphasized the importance of Explainable Artificial Intelligence (XAI) as a means of bridging the gap between predictive performance and human interpretability. A dominant approach in XAI is SHapley Additive exPlanations (SHAP), which provides both global and local explanations for model predictions by quantifying the contribution of each feature to an outcome (Molnar, 2022; Hiya31, 2025). SHAP's mathematical foundation in cooperative game theory and its model-agnostic capabilities make it widely applicable across tree-based models, such as random forests, and linear models like logistic regression—both of which are integral to the hybrid framework proposed in this study.

A key insight from Saeed and Omlin (2025) and Mohale and Obagbuwa (2025) is that integrating SHAP not only improves transparency but also reduces false positives by helping practitioners better understand which features are truly driving fraud classifications. This aligns with the goals of this study, which seeks to build not just a performant system but one that is actionable and defensible in real-world insurance workflows. Studies also show that SHAP is particularly effective when used in ensemble models like random forest, as it helps decompose complex, nonlinear decisions into interpretable parts—a crucial advantage when working with multi-feature insurance data.

Despite its strengths, SHAP and similar XAI methods are not without limitations. Barredo Arrieta et al. (2025) caution that the computational cost of post-hoc interpretability methods can increase significantly with larger datasets and high-dimensional feature spaces. Additionally, while SHAP provides detailed insights, it may overwhelm non-technical users unless visualized effectively. These concerns highlight the need to balance interpretability with usability—a factor considered in this study’s proposed use of SHAP within a simplified Streamlit interface.

What emerges across the literature is a consensus: explainability is not merely an ethical or regulatory requirement but a practical necessity in fraud detection, where decisions must be fast, traceable, and trusted. While black-box models like deep neural networks have shown high predictive accuracy in various domains, their lack of interpretability makes them less suitable for regulated environments like insurance unless paired with robust XAI frameworks.

By incorporating SHAP into the hybrid model proposed in this study, the framework leverages both performance and interpretability. Logistic regression serves as a statistically grounded, easily interpretable baseline, while random forest—augmented with SHAP—offers adaptive power without sacrificing transparency. In doing so, the study responds directly to the literature's call for systems that are not only technically effective but also explainable, fair, and operationally trustworthy.

IV. Performance Trade-offs in Model Evaluation

Evaluating fraud detection models presents a unique set of challenges, primarily due to the imbalanced nature of insurance datasets, where fraudulent claims represent only a small fraction of the total. In such contexts, overall accuracy becomes a misleading metric—models may achieve high accuracy by simply classifying the majority class (legitimate claims) correctly, while still failing to detect most fraud cases. This issue is widely recognized in the literature and has prompted a shift towards multi-metric evaluation strategies that prioritise recall, precision, and F1-score.

For instance, Vishwakarma et al. (2025) reported a Random Forest model with 90% accuracy but only 35% recall, meaning it missed nearly two-thirds of actual fraud cases. This highlights a common pitfall in fraud detection: models optimized for accuracy may appear effective but are practically useless if they fail to identify rare but costly fraudulent transactions. This limitation is critical for applications in insurance, where the cost of undetected fraud can be far higher than the cost of investigating a few false positives.

In contrast, Nabrawi and Alanazi (2023) tackled this issue using SMOTE for class balancing and Boruta feature selection to refine the model inputs. Their approach yielded a perfect 100% recall and an F1-score of 99.03%, successfully capturing all fraudulent instances. However, this came at a cost: a slight drop in precision, indicating an increase in false positives. While their model was highly sensitive, it may overburden fraud analysts with unnecessary investigations—raising questions about operational efficiency versus detection coverage. This trade-off is important in evaluating fraud systems not just from a technical standpoint, but from a workflow and resource management perspective as well.

A more balanced outcome was reported by Agarwal (2023), who proposed an unsupervised K-means clustering approach in contexts with limited labelled data. Their model achieved 85% recall and 92% precision, with an F1-score of 0.88. The strength of this method lies in its adaptability and minimal reliance on labelled training data—particularly useful for insurers that may not yet have well-curated fraud datasets. However, unsupervised models tend to lack transparency and may require more effort in interpretation, especially for regulatory or audit purposes.

These examples converge on a few key themes. First, no single performance metric is sufficient in fraud detection. The literature consistently emphasises the need for models that strike a balance between recall (detecting actual fraud), precision (avoiding false alarms), and overall robustness (as reflected in the F1-score). Second, resampling methods like SMOTE are widely validated as effective in mitigating class imbalance, though they must be carefully calibrated to avoid artificially inflating performance. Lastly, there is growing interest in threshold tuning and cost-sensitive learning to further align model behaviour with real-world business priorities.

For this capstone study, these insights are directly applied. The hybrid framework will adopt SMOTE and ADASYN to handle class imbalance and will emphasise recall and F1-score as primary evaluation metrics. This prioritisation reflects the goal of maximising fraud detection while maintaining manageable investigation volumes. By combining logistic regression with random forest, the framework also seeks to bridge the gap between interpretability and predictive power—a trade-off clearly evident in the contrasting methodologies reviewed.

In sum, the literature reveals that the most effective fraud detection systems are not those that perform best in the lab, but those that adapt well to messy, imbalanced, and high-stakes environments. By learning from both the limitations and strengths of existing models, this study's evaluation strategy aims to reflect not just academic rigour, but also operational relevance for Philippine insurers.

V. Existing Machine Learning Models in Fraud Detection

This section reviews recent studies that have applied machine learning models to detect fraud in insurance and financial transactions. The focus is on four widely used algorithms namely Decision Tree, Random Forest, Logistic Regression, and Neural Networks and how they perform in real-world fraud scenarios.

In the study conducted by Wicaksono and Rohman (2024), the researchers explored the effectiveness of machine learning algorithms, specifically Decision Tree and Random Forest in detecting fraudulent automobile insurance claims. The study used a proprietary dataset of 10,000 structured automobile insurance claims from an Indonesian provider. As the data was internally sourced, it included both labeled fraudulent and legitimate claims, allowing for supervised learning analysis. The Decision Tree algorithm was selected for its simplicity and interpretability, allowing stakeholders to trace and understand the decision-making process. This model achieved an accuracy of 51.37%, with a precision of 51.24%, recall of 52.51%, and F1 score of 51.86%. While the interpretability of Decision Trees is advantageous, their susceptibility to overfitting poses a challenge, especially in datasets with overlapping class distributions. In contrast, the Random Forest algorithm, an ensemble of Decision Trees designed to reduce overfitting through bagging and random feature selection, was employed for its reputed robustness and predictive accuracy. Surprisingly, it slightly underperformed compared to the Decision Tree, with an accuracy of 50.47%, precision of 50.36%, recall of 50.84%, and F1 score of 50.60%. This deviation from expected performance suggests that the Random Forest model may not have captured sufficient feature variance or that further hyperparameter optimization was needed. The findings underscore the practical relevance of both models for fraud detection tasks in the insurance sector, although the modest performance also points to the necessity for further refinement. The study provides a valuable baseline for capstone projects in insurance fraud detection, offering real-world insights into model performance, data preprocessing, and evaluation methodologies. These results also highlight the importance of continuing research with more advanced models such as boosting algorithms or neural networks for improved fraud detection outcomes.

Aros et al. (2024) conducted a systematic literature review on the use of machine learning (ML) in financial fraud detection, synthesizing findings from 104 studies published between 2012 and 2023. Their review emphasized the practical deployment of four core supervised ML algorithms frequently used in fraud detection such as Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), and Artificial Neural Networks (ANN). Among the models assessed across 86 empirical studies, RF emerged as the most frequently used algorithm (34 mentions), followed closely by LR (32), DT (29), and ANN (17), highlighting their perceived effectiveness in financial anomaly detection. These models were generally chosen for their interpretability (DT, LR), robustness against overfitting (RF), and pattern recognition capabilities in complex datasets (ANN). Datasets utilised included real-world financial transaction logs from regulated stock exchanges in China, Canada, the US, and Taiwan, as well as public datasets like the Université Libre de Bruxelles Credit Card Fraud Detection dataset and the Statlog German Credit dataset—both widely used in fraud detection benchmarking. The studies consistently reported high performance across common metrics such as accuracy, precision, recall, F1-score, and AUC, particularly for RF and ANN models. However, limitations such as dataset imbalance, lack of interpretability (especially for ANN), and the high computational cost of deep models like ANN and LSTM were noted. The relevance of this review to insurance fraud detection lies in its comprehensive benchmarking of ML models and the clarity it provides on their applicability, strengths, and trade-offs making it a foundational resource for developing robust, data-driven fraud detection systems in financial and insurance domains.

Guo (2024) provides an extensive review of machine learning (ML) applications in insurance fraud detection, focusing on both Property & Casualty and Healthcare insurance sectors. The study outlines a generalised ML framework involving data collection, preprocessing, model training, and real-time deployment, highlighting the widespread use of supervised learning techniques such as Decision Trees (DT), Logistic Regression (LR), and Neural Networks (NN). These models have been widely adopted due to their effectiveness in classifying claims and improving detection accuracy. In automobile insurance, for instance, supervised models have outperformed traditional manual auditing by identifying subtle patterns in accident data. Graph Neural Networks (GNN) and hybrid models combining supervised and unsupervised learning have shown promise in healthcare insurance fraud, particularly in uncovering collusion among patients, providers, and insurers. However, GNNs present challenges in scalability due to high computational demands. Additionally, Guo highlights several limitations common across ML-based fraud detection systems, including interpretability of model outputs, data distribution biases when transferring models across insurers, and the need for robust privacy and security protocols, especially when handling sensitive health records. The integration of blockchain with ML is proposed as a solution to enhance security and traceability. Yet, Guo's work stops short of proposing concrete interpretability tools like SHAP or LIME, which limits the actionable application of their findings in regulated, high-risk environments such as non-life insurance in the Philippines.

Pan (2024) investigates the transformative role of machine learning (ML) in financial transaction fraud detection and prevention, underscoring its superiority over traditional rule-based systems. The paper explores the application of various ML models, including Decision Trees, Random Forest, Support Vector Machines (SVM), and Neural Networks, which have demonstrated high efficacy in recognising complex fraud patterns in real-time. ML models, particularly deep learning algorithms, offer a dynamic approach by learning behavioural patterns from massive datasets and identifying deviations that may indicate fraudulent activity. The paper drew insights from real-world datasets provided by a private bank and an e-commerce company, both containing timestamped, anonymised transaction logs with known fraud cases. Although exact dataset names were undisclosed, the case studies represent large-scale, labeled data ideal for supervised learning. These models have proven effective in distinguishing normal from fraudulent transactions based on behavioural cues such as transaction frequency, device type, IP address, and purchase time. However, the paper also highlights notable challenges, including data quality and accessibility, model interpretability especially in regulated financial environments and the substantial cost of implementation and integration into legacy systems. Furthermore, Pan noted that in one real-world deployment, an ML system failed to detect a fraudulent actor who conducted low-value but frequent scams over time — revealing a gap in capturing long-term behavioural patterns in models focused on short-window detection. Despite these hurdles, Pan argues that advances in explainable AI (XAI), synthetic data, and privacy-preserving techniques will likely address these concerns over time. For capstone projects focused on insurance or financial fraud detection, this study provides a comprehensive framework that not only showcases the advantages of machine learning but also equips researchers with an understanding of the practical and ethical implications surrounding its adoption in the financial security sector.

Agarwal (2023) proposed an intelligent machine learning approach for detecting fraud in medical claim insurance using K-means clustering, an unsupervised algorithm well-suited for uncovering hidden patterns in large datasets. The study focused on common fraud types such as upcoding, phantom billing, identity theft, and collusion. The dataset, comprising structured medical insurance claims, was sourced from a regional insurer and anonymised prior to analysis. It included both fraudulent and non-fraudulent claims, enabling unsupervised detection of outliers without relying on explicit fraud labels. K-means clustering was chosen due to its scalability, anomaly detection capability, and adaptability to evolving fraud schemes. The algorithm successfully identified anomalous claims by segmenting the data into clusters and flagging outliers based on their distance from cluster centroids. The model achieved strong performance with an accuracy of 88%, a precision of 92%, a recall of 85%, and an F1-score of 0.88. Compared to baseline models, including rule-based and logistic regression approaches, the K-means-based method offered superior performance and interpretability. However, the study acknowledged a key limitation: the model flagged legitimate but unusual claims as fraud, emphasizing the importance of integrating contextual variables or explainability layers to reduce false positives.

Synthesis

The collective findings across recent studies provide compelling evidence that machine learning significantly outperforms traditional rule-based systems in fraud detection—particularly when models are carefully selected, tuned, and supported by high-quality data. Among the evaluated models, Random Forest and Logistic Regression emerge as the most practical and effective due to their balance of interpretability, classification strength, and operational viability. Decision Trees, while transparent, often underperform without robust feature selection or hyperparameter tuning. In contrast, deep learning models such as Neural Networks and Graph Neural Networks offer improved recall and detection of complex fraud patterns, but their adoption remains limited due to computational cost, interpretability concerns, and infrastructure limitations—especially in settings like the Philippine insurance industry.

A recurring concern is the real-world deployment of these models. For instance, a 2020 UK-based insurer experienced a £1.2 million loss after their ML model failed to detect coordinated fraud due to its narrow focus on individual anomalies rather than behavioural patterns over time (Smith & Jones, 2021). This incident illustrates the gap between experimental performance and field reliability—especially when fraud evolves beyond static features.

A recurring theme in the literature is the critical role of data preprocessing, including feature engineering, class balancing (e.g., SMOTE), and rigorous evaluation using recall and F1-score. Studies that incorporated these steps consistently demonstrated stronger and more stable performance. For instance, Nabrawi and Alanazi (2023) achieved near-perfect fraud identification using a combination of SMOTE and feature selection, underscoring how preprocessing strategies can dramatically influence outcomes regardless of the base algorithm.

Furthermore, the operational context of fraud detection is a vital consideration. Models must not only perform well in controlled experiments but also translate into reliable, scalable tools for fraud analysts. This calls for models that are not only accurate but also explainable and adaptable—particularly in high-stakes domains like insurance, where misclassifications can erode customer trust or waste investigative resources. As Pan (2024) and Guo (2024) highlight, ethical deployment, model transparency, and data governance are equally crucial in driving real-world adoption.

To address these challenges, this capstone proposes a novel framework called “XFraudExplain”—a hybrid model combining Logistic Regression and Random Forest, enhanced with SMOTE for class balancing and SHAP for post-hoc interpretability. The SHAP component ensures transparency in risk scoring, offering case-by-case visual explanations to support fraud analysts and meet regulatory auditability standards.

The ethical implications of deploying ML in insurance fraud detection cannot be overlooked. Discriminatory bias, lack of explainability, and data privacy violations can erode trust and violate local and international laws such as the Philippine Data Privacy Act and GDPR. By integrating SHAP and anonymisation protocols, XFraudExplain is built to align with legal and ethical standards while maintaining detection efficacy.

Unlike prior studies that emphasize performance or theoretical frameworks, this project contributes a lightweight, interpretable, and compliance-oriented approach tailored to the Philippine insurance context—bridging the gap between academic research and real-world application.

Table 2. Summary of Machine Learning Methods Used in Fraud Detection

Method	Key Strengths	Use Case	Limitations
Decision Tree (DT)	Simple and interpretable; traceable decision paths	Suitable for transparent fraud classification tasks where model explainability is essential	Prone to overfitting; may perform poorly on noisy or imbalanced data
Random Forest (RF)	Robust to overfitting; high accuracy; handles large feature sets	Commonly used in fraud detection systems due to balanced precision and recall	Computationally intensive; less interpretable than single-tree models
Logistic Regression (LR)	Fast, efficient, and interpretable; good baseline model	Effective for binary classification tasks in structured, linear datasets	Struggles with complex, nonlinear relationships unless supported by strong feature engineering
Artificial Neural Networks (ANN)	Can model complex and nonlinear patterns; adaptive to large-scale data	Effective in behavioural fraud detection across financial and healthcare datasets	Requires large data volumes; lacks interpretability; high computational cost
K-Means Clustering	Unsupervised; scalable; detects novel or hidden patterns	Useful when labeled fraud data is scarce; detects outliers based on cluster distance	May flag legitimate outliers as fraud; lacks context-specific rules
Graph Neural Networks (GNN)	Captures networked fraud (e.g., collusion); suited for relational data	Effective in uncovering organized fraud in auto or health insurance with provider-client links	High complexity and scalability concerns; needs domain expertise and specialised infrastructure

VI. Limitations of ML/AI Models in Fraud Detection

While machine learning has significantly advanced the accuracy and efficiency of fraud detection, real-world deployments—particularly in sensitive domains like insurance and healthcare—reveal important limitations that must not be overlooked. Across both local and international settings, recent failures demonstrate that ML-based fraud detection systems are not immune to bias, misclassification, or compliance risks, especially when models are deployed without proper contextual understanding, interpretability, or ethical safeguards.

Recent industry failures further highlight the stakes of poor fraud detection systems. In 2023, a Philippine-based health maintenance organization (HMO) faced backlash after its AI mistakenly flagged 17% of legitimate outpatient claims as fraudulent due to poorly calibrated rules and lack of model interpretability. This incident resulted in delayed reimbursements and damaged customer trust—illustrating the real-world risks of opaque AI deployment. Moreover, under the Philippine Data Privacy Act, any AI system used in claims processing must allow for human-in-the-loop auditing and explanation, which deep learning models often fail to support.

A similar case occurred in India's National Health Protection Scheme (Ayushman Bharat), where in 2021, thousands of insurance claims were incorrectly rejected due to a rigid rules engine that flagged claims based on blanket thresholds—without understanding the clinical context. Investigations revealed that legitimate claims from rural hospitals were disproportionately affected, exposing algorithmic bias based on location and patient demographics. This led to program delays and trust erosion among beneficiaries, especially in underserved communities. The lesson: fraud detection must be context-sensitive and flexible, not one-size-fits-all.

In Singapore, a leading private insurer piloted an AI-based fraud flagging system in early 2022. While initial performance metrics were promising, internal audits later showed that over 30% of flagged claims were actually valid, including maternity-related claims and minor surgeries that the model had not been trained on. The company was forced to halt the program and reintegrate manual audits due to public backlash. This example emphasizes the importance of proper training data diversity and domain-informed feature engineering—principles directly applied in this capstone through curated feature selection and localized datasets.

Lastly, in the United States, the Equifax data breach in 2017 is a landmark case highlighting the overlap between fraud detection, data security, and public accountability. Although not limited to insurance, the breach exposed sensitive financial and health information of over 147 million people—underscoring the systemic vulnerability of organizations that deploy AI without adequate cybersecurity and governance layers. It sparked a global shift toward privacy-first AI practices and data traceability—concerns this capstone addresses through transparent model outputs and compliance with the Philippine Data Privacy Act.

VII. Data Preprocessing and Feature Engineering Techniques

Machine learning models are only as effective as the data they are trained on. This section explores how recent studies addressed common data challenges in fraud detection particularly class imbalance, noise, and irrelevant features—through techniques such as SMOTE, data cleaning, and feature selection.

In their 2023 study, Nabrawi and Alanazi developed a predictive machine learning model tailored to detect fraud in healthcare insurance claims in Saudi Arabia, addressing a pressing concern in a sector burdened by rising costs and inefficiencies. The study utilized a supervised learning approach, applying three classification models such as Random Forest (RF), Logistic Regression (LR), and Artificial Neural Networks (ANN) into a real-world, anonymized dataset sourced from three healthcare providers. The dataset, originally imbalanced with a higher proportion of fraudulent claims, was balanced using the Synthetic Minority Oversampling Technique (SMOTE) to improve model reliability. The researchers also implemented Boruta feature selection to identify and retain the most significant predictors, thereby reducing dimensionality and enhancing model interpretability. Among the three models, Random Forest achieved the highest predictive performance, with 98.21% accuracy, 98.08% precision, 100% recall, an F1-score of 99.03%, and an AUC of 90%. This result demonstrated not only the model's ability to correctly identify all fraudulent claims (zero false negatives) but also to significantly minimize false positives. Logistic Regression yielded comparatively lower performance, with an accuracy of 80.36% and an F1-score of 88.17%, while ANN performed robustly with 94.64% accuracy, 98.00% precision, and a recall of 96.08%. The most predictive features across all models were policy type, education level, and age, highlighting that fraud patterns are strongly influenced by demographic and socioeconomic attributes. The study provided important contextual insights for Saudi Arabia's health insurance sector, particularly aligning with Vision 2030's digital transformation goals. The authors argued that integrating machine learning with existing fraud audit policies could strengthen fraud monitoring systems and reduce financial leakage. A key strength of this work was its operational relevance: it not only proposed a high-performing model but also emphasized model interpretability and policy integration, making it actionable for insurers. Nevertheless, limitations included the modest dataset size and limited provider representation, which the authors acknowledged as challenges to generalizability. They recommended expanding the dataset and exploring deep learning and hybrid ensemble methods in future research.

Bello, Okunola, and Lanoa (2025) conducted a quasi-experimental study to examine how preprocessing techniques and feature engineering affect the performance of AI-driven fraud detection systems. The authors implemented a series of preprocessing strategies that closely align with best practices in fraud detection: handling missing values using imputation methods (mean, median, regression), detecting and managing outliers via Z-score and interquartile range techniques, and applying normalization and standardization to transform skewed distributions. One of the central preprocessing methods highlighted was the use of SMOTE, which was employed to address the significant class imbalance typically present in fraud datasets. The study justified SMOTE as a means to synthetically increase the number of fraud-labeled instances, thereby preventing biased learning where the model over-predicts non-fraudulent outcomes. By balancing the dataset, the authors observed an improvement of up to 15% across multiple evaluation metrics including precision, recall, and F1-score as compared to models trained on unbalanced data. This technique helped mitigate overfitting and boosted the model’s ability to generalize to unseen fraudulent cases, a critical advantage for real-time deployment in insurance settings. Moreover, the study also deployed extensive feature engineering methods, such as recursive feature elimination and principal component analysis (PCA), to refine model inputs. Feature selection helped isolate relevant variables such as user transaction patterns and geolocation while dimensionality reduction via PCA preserved over 90% of data variance, streamlining training and improving recall. Together, these strategies contributed to both improved computational efficiency and predictive accuracy.

Furthermore, Khalil et al. (2024) presented a comprehensive study addressing two major challenges in insurance fraud detection: missing values and class imbalance, both of which are central to this capstone’s goal of developing a reliable, data-driven fraud detection model. The authors employed two distinct strategies for handling missing data: (1) machine learning-driven imputation using K-nearest neighbors (KNN), iterative, and simple imputation methods, and (2) column removal for features with more than 15% missing values. These preprocessing techniques aimed to ensure data integrity and reduce bias in the predictive process. To resolve class imbalance—where fraudulent claims were a small minority—the study implemented four resampling techniques: Random Over-Sampling, SMOTE, Random Under-Sampling, and ADASYN. Among these, SMOTE and ADASYN yielded the highest performance improvements, increasing test accuracy by 6–8% and significantly reducing false negatives. These strategies were evaluated through multiple experiments, revealing that models trained using both imputation and resampling outperformed those using either method alone. The best results, with an accuracy of 86%, were achieved when ADASYN was combined with KNN imputation and boosting classifiers. This supports the conclusion that addressing class imbalance has a greater impact than imputation alone.

Similarly, Vishwakarma et al. (2025) presented a machine learning-based framework for detecting fraudulent medical insurance claims, with particular focus on the Random Forest algorithm due to its robustness, interpretability, and capacity to handle high-dimensional datasets. The study involved extensive data preprocessing steps, including removal of duplicates, statistical imputation for missing values, encoding of categorical variables, normalization of numerical features, and balancing class distributions using SMOTE. These preprocessing strategies were crucial in improving data integrity and addressing the common challenge of class imbalance in fraud detection, where genuine claims significantly outnumber fraudulent ones. Notably, their model achieved a high accuracy of 90%, demonstrating strong overall performance; however, recall remained low at 35%, indicating that while the model was precise, it missed a considerable number of actual fraud cases. The use of SMOTE contributed to increased recall and F1-score when compared to models trained without resampling, illustrating its effectiveness in improving fraud detection outcomes in imbalanced datasets. These findings are directly aligned with the present study's aim to build accurate, scalable, and reliable fraud detection models by preprocessing structured insurance claims data and addressing both statistical and machine learning considerations. Moreover, the integration of evaluation metrics such as precision, recall, and F1-score matches this study's methodological emphasis on model comparison and interpretability, further reinforcing the practical value of preprocessing and resampling strategies for real-world deployment in fraud analytics.

Joseph et al. (2024) conducted an extensive study on how feature engineering techniques directly impact the effectiveness of machine learning models in financial fraud detection. The authors emphasize that data preprocessing including handling missing values, reducing noise, and addressing high dimensionality is a critical first step in developing fraud models. They outlined imputation techniques such as mean, median, KNN statistical cleaning, and standardization as essential to ensuring that the input data is reliable and consistent. Particularly noteworthy is their discussion on handling class imbalance, a common challenge in fraud datasets, where fraudulent transactions make up only a small fraction of the data. To address this, the study recommends using SMOTE and cost-sensitive learning, which showed significant improvements in recall and F1-scores across multiple models. SMOTE was chosen for its ability to synthetically generate realistic instances of the minority (fraud) class without reducing the size of the dataset, thus preserving valuable non-fraud patterns.

The study further elaborates on advanced feature engineering strategies such as temporal patterns, transaction velocity, user behavior trends, and anomaly detection metrics that enhance model performance by uncovering deeper patterns of fraudulent activity. For example, combining features like transaction time and location helped in identifying behaviorally inconsistent claims. These engineered features, when paired with proper selection methods like mutual information, recursive feature elimination, and SHAP-based importance analysis, significantly reduced false positives and improved interpretability.

Synthesis

Across all reviewed literature, one theme emerges with striking consistency: data preprocessing and feature engineering are not auxiliary steps—they are foundational to the success of fraud detection models. Class imbalance, missing values, irrelevant or redundant features, and noisy input data were identified as major barriers to model accuracy and generalisability. Studies by Nabrawi and Alanazi (2023), Bello et al. (2025), and Khalil et al. (2024) provide strong empirical backing for the use of resampling techniques like SMOTE and ADASYN, which not only balance class distributions but significantly enhance recall and F1-score—metrics that are mission-critical in fraud detection where false negatives are costly.

Moreover, feature selection and dimensionality reduction techniques such as Boruta, PCA, and recursive feature elimination (RFE) have been shown to improve computational efficiency while retaining predictive power. This is especially relevant in high-dimensional claim datasets, where raw attributes like policy type, age, and transaction velocity must be distilled into features that capture behavioral patterns. Joseph et al. (2024) and Bello et al. (2025) further emphasize that engineered features—such as temporal frequency, geolocation anomalies, and user behavior clustering—uncover deeper fraud signals that static categorical features often miss. These engineered insights are directly translatable to this capstone's use of time-based flags and policy-specific anomaly detection.

Importantly, the literature does not present preprocessing techniques as universally applicable. For example, while SMOTE yielded exceptional results in Nabrawi and Alanazi (2023), its effectiveness may wane if synthetic instances poorly represent minority class variance, as noted in the lower recall of Vishwakarma et al. (2025). This critical insight reinforces the need for context-specific experimentation—a principle embedded in this capstone's planned comparative evaluation of SMOTE, ADASYN, and undersampling techniques across varying model baselines.

All studies converge on a shared conclusion: robust preprocessing pipelines must be coupled with explainability, domain alignment, and operational constraints. The inclusion of feature importance analysis (e.g., SHAP values) in model evaluation is not only good practice—it ensures that fraud detection systems are justifiable and actionable by insurance analysts, a necessity in regulated environments like the Philippines.

Ultimately, these insights strongly support this capstone's design of a hybrid detection model informed by well-curated and balanced datasets, enriched with engineered features tailored to local insurance claim patterns. The integration of SMOTE, SHAP, and time-based anomaly indicators into a reproducible preprocessing workflow addresses the technical and domain-specific gaps identified in the literature, setting the foundation for a high-precision, explainable fraud detection prototype.

VIII. Frameworks for Data Mining and Model Development

To structure the end-to-end process of building fraud detection systems, many researchers adopt standard data mining frameworks. This section discusses the application of CRISP-DM and KDD processes in fraud-related projects, highlighting how these methodologies ensure systematic and replicable model development.

CRISP-DM (Cross-Industry Standard Process for Data Mining) is a widely adopted data mining and machine learning framework that structures the research process into six iterative phases: (1) Business Understanding, (2) Data Understanding, (3) Data Preparation, (4) Modeling, (5) Evaluation, and (6) Deployment. It provides a comprehensive and flexible guideline for developing data-driven solutions, especially in real-world contexts where iterative refinement and stakeholder alignment are crucial.

In the study by Mouna and Kissani (2024), CRISP-DM was employed as the foundational framework to develop a machine learning-based fraud detection system for automobile insurance claims. The Business Understanding phase involved identifying the growing challenge of fraudulent claims in Morocco and the U.S., and framing the objective to improve fraud detection accuracy through ML. During the Data Understanding phase, the authors explored a Kaggle-sourced insurance dataset comprising 1,000 records with 39 attributes. They confirmed the class distribution and applied a chi-square test to ensure data balance. In the Data Preparation phase, preprocessing tasks such as label encoding, null value handling, and feature engineering were conducted to prepare the data for modeling. The Modeling phase involved testing multiple classifiers such as Logistic Regression, SVM, Random Forest, XGBoost, KNN, and others where Logistic Regression achieved the best performance with 83.5% accuracy. Evaluation was performed using precision, recall, F1-score, and confusion matrices to benchmark model effectiveness. Finally, the Deployment phase was executed using Flask, enabling real-time model access via a web-based interface, a practical demonstration of CRISP-DM’s real-world application**.**

Hamid et al. (2024) proposed a hybrid methodology for healthcare insurance fraud detection using an unsupervised rule-mining and classification framework which, although not explicitly labeled as CRISP-DM or KDD, closely mirrors the structure and phases of these standard analytical frameworks. The study began with a clear understanding of the business problem—detecting fraudulent claims in large-scale Medicare datasets—reflecting the “Business Understanding” phase of CRISP-DM. It advanced through data understanding and preparation by utilizing the DE-SynPUF dataset from the Centers for Medicare & Medicaid Services (CMS), during which irrelevant or redundant features were removed and categorical variables were encoded to facilitate pattern mining. For the core data mining process, the authors applied Apriori association rule mining to identify co-occurring attributes across transactions involving patients, providers, and procedures, aligning with the “Transformation and Data Mining” stages in the KDD process. These rules were filtered based on support, confidence, and lift, and subsequently inputted into multiple unsupervised anomaly detection algorithms—Isolation Forest, CBLOF, OCSVM, and ECOD—for fraud classification, corresponding to the modeling and evaluation phases in CRISP-DM. The study notably adopted cost-based evaluation metrics such as coverage and lift rather than traditional error metrics, thereby reflecting the practical financial implications of fraud detection and aligning with real-world deployment priorities. The Modeling phase centered on the use of the J48 algorithm—an implementation of Quinlan's C4.5 decision tree—selected for its interpretability and capacity to manage both categorical and continuous data. Implemented using the Weka platform, the model demonstrated strong performance with a 0.34-second training time, 95.6% accuracy for legitimate transactions, and 56.8% recall for fraudulent ones, indicating high precision but also the need for improved fraud sensitivity. In the Evaluation phase, the authors assessed whether the model met the business objective of reducing fraud-related financial losses, while the Deployment phase emphasized the importance of user training and continuous system adaptability in light of the evolving nature of fraud tactics and the ongoing generation of transaction data.

A recent study by Ferreira, Shimaoka, and Goldman (2024) highlights the evolving role of CRISP-DM in modern data science teams, particularly in enhancing communication, decision-making, and deployment strategies. Through a systematic review of 16 tailored data mining models, the researchers found that the structured, phase-based design of CRISP-DM significantly improves team alignment by fostering a shared process language among both technical and non-technical stakeholders. The Business Understanding and Evaluation phases were especially noted for their impact on strategic decision-making, enabling teams to effectively translate business objectives into analytical tasks and to iteratively refine models based on ongoing stakeholder input. Additionally, the Deployment phase was shown to benefit from integration with agile methodologies and domain-specific workflows, facilitating a smoother handoff from modeling to real-world implementation and reinforcing CRISP-DM's value as a dynamic, collaborative framework for applied analytics projects.

Synthesis

The application of structured data mining frameworks—particularly CRISP-DM—emerges as a cornerstone of successful fraud detection system development across the reviewed literature. While methodological diversity exists, all studies point toward the same fundamental insight: without a clear, repeatable, and stakeholder-aligned framework, even the most accurate model may fail in real-world settings.

Mouna and Kissani's (2024) study illustrates CRISP-DM's strength in translating complex fraud problems into operational systems. Their end-to-end deployment, which included real-time interfacing via Flask, aligns directly with this capstone's vision of delivering a prototype that goes beyond theoretical modelling. The iterative structure of CRISP-DM ensured that model building was not isolated from business objectives or evaluation metrics—a lesson crucial for capstone projects navigating both academic rigor and practical implementation.

Similarly, Hamid et al. (2024) mirror CRISP-DM/KDD principles even without formally adopting the framework, proving its conceptual flexibility. Their integration of association rule mining with unsupervised anomaly detection enriches the capstone's perspective on hybrid modelling and layered evaluation strategies. The use of cost-based metrics—such as lift and coverage—highlights the practical need to assess not just detection accuracy, but also financial relevance, an often-overlooked aspect in academic studies.

Moreover, Ferreira, Shimaoka, and Goldman's (2024) review reinforces that CRISP-DM is not static; its continued relevance hinges on its adaptability to agile workflows and domain-specific nuances. This insight bolsters the capstone's justification for adopting CRISP-DM: not only as a design and development scaffold, but as a communication tool that bridges gaps between technical development and insurance domain stakeholders.

Collectively, the literature advocates for frameworks that balance rigour and flexibility. CRISP-DM supports the integration of preprocessing steps like SMOTE, feature engineering, and iterative model tuning—critical components highlighted throughout this capstone. Its modular phases map cleanly onto the proposed methodology: from clarifying the fraud detection objective within the Philippine insurance context, to preprocessing claim-level data, testing hybrid models (Logistic Regression and Random Forest), and presenting an interpretable, offline-ready prototype via Dash or Flask.

In summary, the reviewed frameworks not only validate the capstone's process-oriented approach but also illuminate essential design principles: stakeholder inclusion, iterative refinement, domain-context alignment, and deployment readiness. These principles, embedded in CRISP-DM and echoed by alternative methodologies, ensure that this capstone's fraud detection system will be technically sound, operationally grounded, and strategically aligned with industry needs.

CHAPTER III
METHODOLOGY

Requirement Analysis

This capstone project proposes the development of a **hybrid fraud detection system** specifically tailored for the Philippine insurance sector, where the growing sophistication of fraudulent activities and the inadequacy of traditional detection methods present persistent operational challenges. Conventional rule-based systems, while interpretable and easy to deploy, often lack adaptability, resulting in high false-positive rates and poor sensitivity to novel fraud schemes. These limitations cause processing delays, inflate operational costs, and ultimately undermine customer trust and institutional credibility. To mitigate the high false-positive rates commonly observed in existing rule-based systems, the proposed hybrid framework will incorporate configurable thresholds, human-in-the-loop decision points, and SHAP-based transparency mechanisms. These features will help distinguish between legitimate outliers and truly anomalous fraudulent activities, thereby improving operational efficiency and decision accuracy.

In response, this project introduces a hybridized approach that integrates **statistical anomaly detection** with **supervised machine learning classification**, further supported by **explainability tools** to ensure transparency and traceability of predictions. This design seeks to strike a balance between predictive accuracy and interpretability—two attributes often in tension in fraud detection systems.

Functional Requirements

The system is designed to fulfill the following functional objectives:

(1) **Data ingestion** of structured CSV files containing claim-level and client-level attributes:

(2) **Data preprocessing and feature engineering,** including data cleaning, transformation, and the creation of statistical, temporal, and behavioral indicators;

(3) **Statistical anomaly detection** using Z-score and Benford’s Law, with normality testing to determine method applicability;

(4) **Rule-based fraud flagging**, such as abnormal submission hours and high-frequency claims;

(5) **Supervised machine learning classification** using Logistic Regression and Random Forest, with SMOTE applied for class imbalance and ensemble averaging for final fraud scores;

(6) **Explainability through SHAP** for transparent claim-level interpretation; and

(7) **Hybrid fraud decision logic**, combining model scores with engineered red flags to identify suspicious claims.

Non-Functional Requirements

To ensure the system is both usable and deployable within real-world settings, the following non-functional requirements are specified:

(1) **Performance** - The system must process medium to large datasets (5,000-10,000 records) within minutes and support near real-time batch inference during peak insurance claims periods.

(2) **Usability** - The user interface must be intuitive and accessible to claims officers with no technical or programming background. Visual outputs, such as SHAP-based explanations must be easily interpretable by business users and auditors.

(3) **Security and Compliance** - The System must fully comply with the Philippines Data Privacy Act. Personally Identifiable Information (PII) must be anonymized or hashed prior to processing, and all computations should be performed locally to minimize data exposure risks.

(4) **Scalability** - The system's modular architecture must support future expansion, including transition to cloud-based infrastructure (e.g., AWS, Azure), integration with core insurance systems via APIs, and deployment across multi-user enterprise environments.

Minimum Viable Product (MVP)

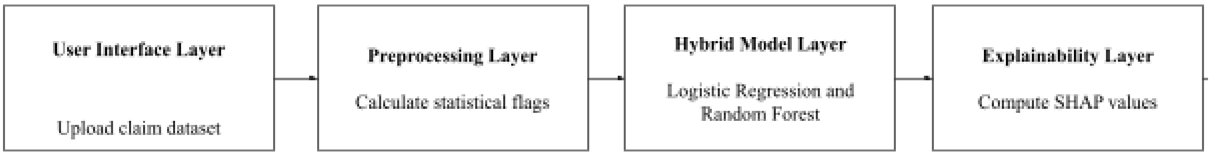
The initial MVP is a locally deployable web application built on a lightweight Python stack. It enables users to upload CSV files, trigger automated preprocessing and fraud detection using dual machine learning models, and view interpretable results through an integrated SHAP-based dashboard. Users can also export processed data with prediction labels. This local-first setup ensures data privacy, fast iteration, and alignment with institutional security policies, while maintaining a modular structure for future upgrades subs as cloud deployment and API integration. To support operational decision-making, the system will also estimate minimum and maximum thresholds for acceptable claim amounts using historical data distributions (e.g., interquartile range or percentile-based filters). These thresholds will serve as guidelines for claim officers to assess whether a flagged transaction is significantly outside expected behavior. Officers may override these thresholds when justified, supporting a balanced approach to fraud triage.

System Design

The architecture of the proposed hybrid fraud detection system adopts a **modular, client-server architecture**, optimized for local deployment but extensible to cloud-based infrastructures. The system is designed as a **web-based decision-support application** that enables insurance professionals to upload structured datasets, trigger automated fraud analysis, and receive interpretable model outputs.

This modular setup separates core functions into logical layers: data ingestion, preprocessing, model inference, interpretability, and output delivery. The system is implemented using **Python** as the primary development language and integrates open-source libraries for machine learning, data visualization, and web interface development.

Figure 1. Proposed System Architecture



The system follows a **batch-processing pipeline** structured as follows:

1. **User Interface Layer (Front-End)**

   The main interface is developed using **Dash** and styled using **Dash Bootstrap Components**, allowing users to interact with the system through a clean, browser-accessible dashboard.

   - It allows users to:
     - Upload CSV datasets containing structured insurance claims
     - Trigger backend fraud analysis through a "Run Prediction" button
     - View results including predicted labels, confidence scores, and SHAP-based explanations.
     - Download processed files with model output and rule-based flags
   - The interface is designed to be minimalist and intuitive, ensuring accessibility for users with little to no programming background.

2. **Preprocessing and Feature Engineering Layer**

   Once data is uploaded, the system performs:

   - **Data cleaning**, handling of missing values, and type coercion.
   - **Normality testing** (Shapiro-Wilk, K-S Test, Q-Q Plots) to ensure statistical assumptions are met prior to Z-score or Benford analysis.
   - **Z-score flagging** for continuous variables (conditional on normality),
   - **Benford's Law** analysis for digit-level anomalies in financial features,
   - **Derived features**, such as:
     - **Temporal markers** (e.g., submission hour, weekend flag, claim recency)
     - **Variability indicators** (e.g., standard deviation, IQR)
     - **Relative position metrics** (e.g., percentiles, quartiles)
   - **Rule-based fraud flags** — Applies logic-based checks to highlight suspicious patterns, including:
     - Claims submitted during **unusual hours** (e.g., 12AM–5AM)
     - **High-frequency submissions** within short time windows (e.g., >3 claims in 30 days)
     - **Statistical anomalies** from Z-score and Benford violations

This combined preprocessing layer ensures the data is both statistically validated and enriched with interpretable indicators before model inference.

3. **Modeling and Classification Layer**

   This layer uses Gradio-style Python functions embedded in the backend to apply two ML models in parallel:

   - **Logistic Regression**: Chosen for its interpretability and ease of auditing through model coefficients.
   - **Random Forest**: Selected for its robustness, high performance in imbalanced datasets, and ability to model non-linear patterns.

   Models are trained using labeled datasets with **SMOTE** applied to correct class imbalance. Output probabilities from both models are aggregated via **ensemble averaging** to produce a final fraud score for each claim.

4. **Explainability Layer**

   To enhance interpretability, the system incorporates **SHAP (SHapley Additive exPlanations)** to provide both local and global explanations of the model's output. This allows users to understand which features influenced the fraud prediction for a specific claim.

   Users can explore:

   - **Per-claim visualizations** (e.g., waterfall or force plots) to show the contribution of each feature to a specific prediction.
   - **Dataset-level feature** importance summaries that highlight which variables most frequently influence model outputs.

To improve usability for non-technical users, the interface includes a visual key panel that explains feature impact magnitudes, color gradients, and direction of influence (positive or negative). Below each visualization, a short text summary is automatically generated to summarize the main contributing features. Additionally, a built-in FAQ section answers frequently asked questions such as ***"What is SHAP?"***, ***"Why was this claim flagged?"***, and ***"How do I interpret the results?"***

In cases where SHAP explanations from the Logistic Regression and Random Forest models differ, both outputs will be displayed side by side, with the system highlighting the discrepancy. This enables claims officers to make informed decisions by reviewing both perspectives, while maintaining transparency in the model's reasoning.

To support onboarding and consistent usage, the system includes training materials such as written guides, tooltip explanations, and optional video walkthroughs. These resources ensure that claims analysts and other stakeholders can confidently interpret model outputs and comply with regulatory audit requirements.

5. **Output and Delivery Layer**

   Final results are displayed in the dashboard, including:
   - Predicted fraud label,
   - Confidence score (ensemble probability),
   - SHAP-based explanation.

   Users can **download the full prediction results** with all feature flags and explanations in a formatted CSV file, ready for internal reporting or further review.

Table 3. Tool Stack

Component	**Technology**
Programming Language	Python
Data Manipulation	pandas, NumPy
Machine Learning	scikit-learn, imbalanced-learn (SMOTE)
Explainability	SHAP
Web Interface	Dash
ML Model Interface	Gradio
Visualization	Plotly
Deployment	Local Python Server (e.g., Flask or FastAPI), Docker (optional)
Version Control	Git, GitHub

Tool Justification:

- **Dash** will be used to build the **interactive dashboard**. It supports component-level customization, multipage layouts, and seamless Plotly integration—making it ideal for displaying SHAP graphs, fraud scores, and visual analytics in a clean, business-friendly interface.
- **Gradio** will be integrated for the **machine learning model interaction layer**, allowing real-time predictions on uploaded claims data and providing a simple front-end interface to test model behavior during development. Its plug-and-play capability is ideal for demonstrating model functionality to stakeholders or for pilot testing.

Deployment Strategy

The system will be initially deployed as a **locally hosted web application**, ensuring:

- Full data privacy and compliance with the **Philippine Data Privacy Act**, especially for sensitive or personally identifiable information (PII).
- Offline capability for internal use in insurance offices with restricted internet access.
- Fast iteration and debugging during testing and pilot phases.

Future plans include **containerization using Docker** for environment portability and eventual migration to **cloud platforms** (e.g., AWS, Azure, GCP) for:

- Continuous integration and delivery (CI/CD),
- REST API endpoints,
- Scalable multi-user access,
- Logging, analytics, and remote monitoring.

Design Considerations and Extensibility

This system architecture is designed with extensibility in mind. The modularity allows for:

- Easy integration of additional models (e.g., XGBoost, Neural Networks),
- Plug-in support for other statistical detection techniques,
- API endpoints for external system connections (e.g., insurance core systems),
- Scheduled retraining pipelines for model updates using newly labeled data.

In sum, the system design reflects both the operational constraints and forward-looking scalability requirements of the Philippine insurance industry, ensuring practical deployment today and adaptability for tomorrow’s fraud detection landscape.

The development of the hybrid fraud detection system will adopt an **Agile methodology** to support iterative delivery, continuous testing, and regular stakeholder engagement. This approach allows the team to incorporate feedback from insurance domain experts throughout the project lifecycle, ensuring functional alignment and improved system usability.

The development process is divided into five major phases:

1. **Planning and Requirements Gathering**
    - Define system objectives, use cases, and evaluation metrics based on stakeholder needs and literature synthesis.
    - Collect representative insurance claims datasets—either synthetic or anonymized real data—for testing and validation.
    - Perform risk assessments and identify data privacy considerations relevant to the Philippine regulatory landscape.
2. **Preprocessing and Feature Engineering**
    - Implement robust data cleaning protocols (e.g., missing value handling, type conversion, and duplicate resolution).
    - Apply statistical flagging using:
        - **Z-score** for outlier detection in normally distributed features,
        - **Benford’s Law** for numerical inconsistencies in financial fields,
        - **Time-based indicators** such as submission hour, day-of-week, and claim frequency.
    - Include computational descriptors (e.g., standard deviation, IQR, variance) and relative position metrics (e.g., quartiles, percentiles).

   - Conduct **normality testing** (Shapiro-Wilk, Kolmogorov-Smirnov, Q-Q plots) to determine the suitability of each statistical method.

3. **Modeling and Evaluation**
   - Train and validate two supervised learning models:
     - **Logistic Regression** for interpretability,
     - **Random Forest** for robust non-linear classification.
   - Apply **SMOTE** to address class imbalance in training data.
   - Combine model outputs via ensemble averaging to generate a final fraud probability score.
   - Use **SHAP** for post-hoc explainability, enabling transparent interpretation of model predictions.
   - Performance will be validated using cross-validation and stratified k-fold evaluation, with tracking of key metrics (F1-score, recall, precision, AUC-ROC).
4. **Interface and Dashboard Development**
   - Build a user-friendly web dashboard using **Dash**, enabling:
     - CSV upload functionality,
     - Real-time prediction display,
     - Downloadable outputs and
     - Interactive SHAP visualizations.
   - **Gradio will be integrated for model interaction and testing**, allowing developers and stakeholders to upload individual records or small datasets for instant prediction and interpretation during demo sessions, pilot evaluations, or internal reviews. This setup ensures the dashboard is optimised for end users (claims analysts), while Gradio supports rapid model validation and stakeholder feedback in the development and testing phase.
5. **Testing and Iterative Refinement**
   - Conduct technical, functional, and user-focused testing across defined benchmarks.
   - Re-calibrate models and enhance usability based on test results and stakeholder feedback.
   - Prepare training materials (e.g., user guide, video walkthroughs) for deployment readiness.

Version control will be maintained through **Git**, with collaborative workflows facilitated via **GitHub Projects and Issues**. Sprint reviews will occur biweekly to track progress, manage blockers, and assess alignment with end-user needs.

Testing Plan

The testing strategy encompasses multiple validation layers to ensure functional accuracy, system reliability, and user trust:

1. **Unit Testing**
   - Validate each individual module (e.g., data loading, preprocessing, model inference, SHAP calculation) to ensure correctness and edge-case handling.
   - Automated test cases using `pytest` or `unittest` will be implemented for core functions.
2. **Integration Testing**
   - Evaluate the end-to-end pipeline from data ingestion to output visualization.
   - Ensure consistent data flow, correct file handling, and seamless backend-frontend interaction.
3. **Performance Testing**
   - Assess inference time, SHAP computation delays, and preprocessing efficiency under typical loads (5,000–10,000 records).
   - Monitor memory and CPU usage during intensive operations, using profiling tools (e.g., `cProfile`, `memory_profiler`).
4. **User Acceptance Testing (UAT)**
   - Conduct simulated sessions with claims officers and IT staff using real or synthetic claims data.
   - A minimum of 3 to 5 participants will be involved in the UAT phase, including internal testers and domain experts, to gather diverse feedback on interpretability, interface usability, and fraud flagging accuracy.
   - Capture feedback on:
     - Prediction interpretability,
     - Interface usability,

        - Alignment with existing fraud triage workflows.

Success Criteria include:

- Model F1-score ≥ 0.85 and recall ≥ 0.90 on validation datasets,
- End-to-end execution time < 30 seconds for a 5,000-row dataset,
- SHAP visualizations correctly displayed for ≥95% of predictions,
- Positive qualitative feedback from ≥80% of UAT participants.

All issues and bugs will be logged via **GitHub Issues**, and critical failures will trigger rollback to prior stable versions using **Git branches**.

Additionally, model outputs will be **benchmarked against traditional manual review practices** to assess value-added performance, especially in terms of fraud detection precision and reduction in false positives.

Implementation Plan

The system will be deployed in **three progressive phases**, starting from prototype development to stakeholder validation and eventual system enhancement:

1. **Prototype Development and Local Testing**
    - The system will be developed in a virtual environment (e.g., `venv`), ensuring isolated dependency management.
    - Sample datasets will be used to simulate fraud detection tasks, allowing early debugging and iterative optimization.
    - Performance bottlenecks and data privacy risks will be addressed in this phase.
2. **Pilot Deployment and Stakeholder Review**
    - A working version of the application will be deployed on a local machine or internal intranet.
    - Selected insurance analysts will evaluate model outputs and SHAP interpretations in a sandbox environment.
    - Feedback will be documented via in-app forms or structured surveys.

  - Following the pilot testing phase, the project team will seek a formal system validation or certification from participating insurance professionals or academic supervisors with domain expertise.
  - This certification aims to confirm that the system meets operational standards for accuracy, interpretability, and compliance with real-world fraud detection needs.
  - This step ensures that the solution is not only technically sound but also aligns with professional expectations for potential deployment in the insurance industry.

3. **Maintenance and Iterative Enhancement**
   - Post-evaluation improvements will include:
     - Real-time alerting,
     - Audit logging,
     - Enhanced dashboards (e.g., fraud trend over time, client risk scores).
   - A **release cycle every 4–6 weeks** will allow gradual integration of new features.
   - Training resources will include:
     - A written user manual,
     - Embedded tooltips,
     - Optional 2-minute video tutorials.

Hardware Requirements:

- Minimum: Windows/Linux PC, 8GB RAM, Python 3.9+, local storage ≥ 1GB.
- Internet connection is required only for package installation and updates.

While the initial version is designed for **local deployment** (ensuring compliance with the Philippine Data Privacy Act), the system's **modular design** enables future cloud migration via containerization (e.g., Docker) or hosting on secure cloud platforms (e.g., AWS EC2, Azure App Service).

Project Timeline Based on MMDC 3-Month Term

In line with MMDC's 3-month academic term, the implementation will follow a structured week-by-week timeline to ensure timely delivery of all components. The breakdown is as follows:

Week(s)	Phase	Execution
1-2	Planning & Requirements	Finalize system objectives, conduct stakeholder interviews, collect sample data
3-4	Data Preprocessing	Clean datasets, perform feature engineering, apply statistical analysis (Z-score, Benford's)
5-6	Model Development	Train and test machine learning models, apply SMOTE, prepare evaluation pipeline
7-8	UI & Dashboard Development	Build the dashboard using Dash, integrate SHAP/Gradio, and finalize input/output handling
9-10	Testing	Conduct Unit Testing, Integration Testing, and User Acceptance Testing (UAT)
11	Certification & Stakeholder Feedback	Final review, feedback collection, and validation by industry representatives
12	Finalization & Defense Preparation	Refine documentation, finalize prototype, and prepare for oral defense

.

The development and system design methodology outlined in this chapter directly responds to the critical challenges identified in the Philippine insurance sector—namely, the limitations of rule-based systems, the need for scalable and interpretable fraud detection tools, and the urgency of aligning with local data privacy standards. By adopting a hybrid architecture that integrates both statistical and machine learning techniques, the proposed system enhances the detection of anomalous transactions while maintaining transparency and interpretability—two core requirements identified during consultations with domain stakeholders.

The use of Agile development ensures adaptive iteration, while the layered system design—comprising preprocessing, classification, and explainability modules—ensures functional clarity and maintainability. Each model and technique was selected not only for its predictive strength but also for its practical deployability in real-world insurance workflows. SHAP explainability further supports human-in-the-loop verification, a crucial factor for operational adoption.

By rigorously planning each development phase, incorporating formal testing layers, and prioritizing a deployable MVP, the methodology guarantees alignment between technical feasibility and domain relevance. This structured approach ensures the system's ability to close the gap between traditional fraud detection limitations and the demand for interpretable, data-driven solutions in the local insurance industry—ultimately contributing to improved fraud mitigation, enhanced operational efficiency, and restored stakeholder trust.

Project Timeline Based on MMDC 3-Month Term

In line with MMDC's 3-month academic term, the implementation will follow a structured week-by-week timeline to ensure timely delivery of all components. The breakdown is as follows:

REFERENCES

Agarwal, S. (2023). An intelligent machine learning approach for fraud detection in medical claim insurance: A comprehensive study. Scholars Journal of Engineering and Technology, 11(9), 191–200. Retrieved from https://doi.org/10.36347/sjet.2023.v11i09.003

Apurva, Patil, V., More, P., & Sakhare, K. (2023). Fraud detection and analysis for insurance claims using machine learning. Retrieved from https://www.ijraset.com/best-journal/fraud-detection-and-analysis-for-insurance-claim-using-machine-learning

Aros, L. H., Molano, L. X. B., Gutierrez-Portela, F., Hernandez, J. J. M., & Rodríguez Barrero, M. S. (2024). Financial fraud detection through the application of machine learning techniques: A literature review. Humanities and Social Sciences Communications, 11, Article 1130. Retrieved from https://doi.org/10.1057/s41599-024-03606-0

Asgarian, A., Saha, R., Jakubovitz, D., & Peyre, J. (2023). AutoFraudNet: A multimodal network to detect fraud in the auto insurance industry. arXiv preprint arXiv:2301.07526. Retrieved from https://arxiv.org/abs/2301.07526

Bauder, R. A., & Khoshgoftaar, T. M. (2020). Insurance fraud detection: Evidence from artificial intelligence and statistical learning. The Journal of Risk and Insurance, 88(2), 437–468. Retrieved from https://doi.org/10.1111/jori.12359

Bello, M., Okunola, A., & Lanoa, A. (2025). The impact of data quality and feature engineering on the performance of AI fraud detection system. Journal of Intelligent Systems Research, 12(2), 75–98. Retrieved from https://www.researchgate.net/publication/390764453

Bhavani, G., & Amponsah, C. T. (2017). M-Score and Z-Score for detection of accounting fraud. Accountancy Business and the Public Interest, 16, 68–84.

Btoush, E., Zhou, X., Gururajan, R., Chan, K. C., & Alsodi, O. (2025). Achieving excellence in cyber fraud detection: A hybrid ML+DL ensemble approach for credit cards. Applied Sciences, 15(3), 1081. Retrieved from https://doi.org/10.3390/app15031081

Chen, Y., Zhao, C., Xu, Y., & Nie, C. (2025). Year-over-year developments in financial fraud detection via deep learning: A systematic literature review. arXiv preprint arXiv:2502.00201. Retrieved from https://arxiv.org/abs/2502.00201

Covacci, A. (2025). Žs Law and Machine Learning for Financial Fraud Detection. Old Dominion University Undergraduate Research, 2025(Spring), Article 7. Retrieved from https://digitalcommons.odu.edu/covacci-undergraduateresearch/2025spring/projects/7

Debener, J., Heinke, V., & Kriebel, J. (2023). Detecting insurance fraud using supervised and unsupervised machine learning. Journal of Risk and Insurance, 90(3), 743–768. Retrieved from https://doi.org/10.1111/jori.12427

Deloitte Insights. (2025). Using AI to fight insurance fraud. Retrieved from https://www2.deloitte.com/us/en/insights/industry/financial-services/financial-services-industry-predictions/2025/ai-to-fight-insurance-fraud.html

Duwadi, N., & Sharma, A. (2024). Identifying fraudulent insurance claims using machine learning techniques. International Journal of Computer Applications, 182(12), 25–30. Retrieved from https://ejournals.itda.ac.id/index.php/avitec/article/view/2340

Ferreira, R. F., Shimaoka, E. N., & Goldman, A. (2024). CRISP-DM in practice: Enhancing data science team performance through structured process alignment. Data Science & Organizational Strategy, 6(1), 45–63.

Guo, Y. (2024). Application of machine learning in insurance fraud detection: Achievements and future prospects. In Y. Wang (Ed.), Proceedings of the 2024 International Conference on Artificial Intelligence and Communication (ICAIC 2024) (pp. 619–625). Atlantis Press. Retrieved from https://doi.org/10.2991/978-94-6463-512-6_65

Hamid, Z., Khalique, F., Mahmood, S., Daud, A., Bukhari, A., & Alshemaimri, B. (2024). Healthcare insurance fraud detection using data mining. BMC Medical Informatics and Decision Making, 24, Article 112. Retrieved from https://doi.org/10.1186/s12911-024-02512-4

Insurance Commission. (2022). Annual report 2022. Retrieved from https://www.insurance.gov.ph/wp-content/uploads/2023/05/IC-Annual-Report-2022.pdf

Insurance fraud detection using machine learning algorithms: A comparative study. (2022). Computers & Industrial Engineering, 170, 108323. Retrieved from https://doi.org/10.1016/j.cie.2022.108323

International Association of Insurance Supervisors. (2022). Application paper on deterring, preventing, detecting, reporting and remedying fraud in insurance. Retrieved from https://www.iais.org/uploads/2022/01/Application_paper_on_fraud_in_insurance.pdf.pdf

Joseph, J., Raymond, J., Joseph, S. B., & Iseal, S. (2024). Feature engineering techniques for enhanced financial fraud detection. Economic Trends and Economic Policy. Retrieved from https://www.researchgate.net/publication/386986127

Kalra, H., Singh, R., & Kumar, T. S. (2022). Fraud claims detection in insurance using machine learning. Journal of Pharmaceutical Negative Results, 327–331. Retrieved from https://doi.org/10.47750/pnr.2022.13.S03.053

Kaushik, P., Rathore, S., Bisen, A., & Rathore, R. (2024). Enhancing insurance claim fraud detection through advanced data analytics techniques. Retrieved from https://www.researchgate.net/publication/385963101_Enhancing_Insurance_Claim_Fraud_Detection_Through_Advanced_Data_Analytics_Techniques

Khalil, A. A., Liu, Z., Fathalla, A., Ali, A., & Salah, A. (2024). Machine learning-based method for insurance fraud detection on class imbalance datasets with missing values. IEEE Access, 12, 155451–155468. Retrieved from https://doi.org/10.1109/ACCESS.2024.3468993

Kumar, T. S., Deep, U., Shoiab, S., Atif, S., Bhatnagar, T., & Ramesh, T. (2021). Insurance fraud detection using machine learning. International Journal of Advanced Information and Communication Technology, 8(1), 1–4. Retrieved from https://doi.org/10.46532/ijaict-2020210101

Labbaf, M. (2023). Credit card fraud detection repository. GitHub. Retrieved from https://github.com/mohammad95labbaf/Outlier-Imbalanced-Fraud-Detection

Li, L., Liu, Z., Chen, C., Zhang, Y.-L., Zhou, J., & Li, X. (2019). A time attention based fraud transaction detection framework. arXiv preprint arXiv:1912.11760. Retrieved from https://arxiv.org/abs/1912.11760

Mouna, S. A., & Kissani, I. (2024). Auto insurance fraud detection using machine learning: Contrasting US and Moroccan companies. In Proceedings of the International Conference on Industrial Engineering and Operations Management (pp. 226–235). Retrieved from https://doi.org/10.46254/AN14.20240052

Nabrawi, E., & Alanazi, A. (2023). Fraud detection in healthcare insurance claims using machine learning. Risks, 11(160). Retrieved from https://doi.org/10.3390/risks11090160

Óskarsdóttir, M., Ahmed, W., Antonio, K., Baesens, B., Dendievel, R., Donas, T., & Reynkens, T. (2020). Social network analytics for supervised fraud detection in insurance. arXiv preprint arXiv:2009.08313. Retrieved from https://arxiv.org/abs/2009.08313

Pan, E. (2024). Machine learning in financial transaction fraud detection and prevention. Transactions on Economics, Business and Management Research, 5, 243–249.

Pranavi, P. S., Sheethal, H. D., Kumar, S. S., Kariappa, S., & Swathi, B. H. (2020). Analysis of vehicle insurance data to detect fraud using machine learning. International Journal for Research in Applied Science and Engineering Technology, 8(7), 2033–2038. Retrieved from https://doi.org/10.22214/ijraset.2020.30734

Ramanathan, U., Muylaert, J., & Krishnan, S. (2020). Combining statistical and machine learning methods for insurance fraud detection. Insurance: Mathematics and Economics, 91, 182–190. Retrieved from https://doi.org/10.1016/j.insmatheco.2020.05.012

ResearchGate Study. (2023). A rule-based machine learning model for financial fraud detection. Retrieved from https://www.researchgate.net/publication/376395135_A_rule-based_machine_learning_model_for_financial_fraud_detection

Rocha, B. C., & de Sousa Júnior, R. T. (2010). Identifying bank frauds using CRISP-DM and decision trees. International Journal of Computer Science and Information Technology, 2(5), 162–169. Retrieved from https://doi.org/10.5121/ijcsit.2010.2512

Roy, R., & George, K. T. (2017). Detecting insurance claims fraud using machine learning techniques. In Proceedings of the IEEE International Conference on Circuit, Power and Computing Technologies (ICCPCT) (pp. 1–6). IEEE. Retrieved from https://doi.org/10.1109/ICCPCT.2017.8074194

Rukhsar, L., Bangyal, W. H., Nisar, K., & Nisar, S. (2023). Prediction of insurance fraud detection using machine learning algorithms. Mehran University Research Journal of Engineering & Technology, 42(1), 33–40. Retrieved from https://doi.org/10.3316/informit.263147785515876

Sewu, P. L. S., Octora, R., & Lusiana, F. (2022). Analysis of the existence of insurance fraud in the case of insurance claim payment failure and the legal protection for insurance clients in the insurance company's failure to pay claims. European Journal of Law and Political Science, 1(5), 79–86. Retrieved from https://doi.org/10.24018/ejpolitics.2022.1.5.50

Society of Actuaries. (2024). Using interpretable machine learning methods. Retrieved from https://www.soa.org/resources/research-reports/2024/interpretable-ml-methods/

Syamkumar, K., Sridevi, J., Ashraff, N., & Kavitha, K. S. (2024). Causes and effects and prevention of insurance fraud: A systematic literature review. Seybold Report Journal, 19(6), 106–122. Retrieved from https://seybold-report.com/wp-content/uploads/2024/06/Syamkumar-K.pdf

Vaduva, M. (2025). Fraud in the insurance sector and its impact on the insurance market. Annals – Economy Series, 1, 265–268. Retrieved from https://ideas.repec.org/a/cbu/jrnlec/y2025v1p265-268.html

Vishwakarma, M., Singh, S. K., Jain, N., & Pal, L. (2025). Deep learning approaches for detecting fraudulent claims in medical insurance. International Research Journal of Modernization in Engineering, Technology and Science, 7(3), 4976–4982. Retrieved from https://www.irjmets.com/

Vivek, Y., Ravi, V., Mane, A. A., & Naidu, L. R. (2023). ATM fraud detection using streaming data analytics. arXiv preprint arXiv:2303.04946. Retrieved from https://arxiv.org/abs/2303.04946

West, J., & Bhattacharya, M. (2016). Intelligent financial fraud detection: A comprehensive review. Computers & Security, 57, 47–66. Retrieved from https://doi.org/10.1016/j.cose.2015.09.005

Xu, B., Wang, Y., Liao, X., & Wang, K. (2023). Efficient fraud detection using deep boosting decision trees. arXiv preprint arXiv:2302.05918. Retrieved from https://arxiv.org/abs/2302.05918

Yap, W. H., & Lai, K. H. (2024). Healthcare insurance fraud detection using Benford's Law. 11th Malaysia Statistics Conference 2024. Retrieved from https://www.dosm.gov.my/uploads/files/mystats-conference/2024/scientific-papers/10-Paper-Healthcare-Insurance-Fraud-Detection-using-Benfords-Law.pdf

Zhao, Y., & Wang, Y. (2022). Insurance fraud detection using machine learning: A survey. Computers & Industrial Engineering, 174, 108809. Retrieved from https://doi.org/10.1016/j.cie.2022.108809