

PayGuard AI: An Explainable Hybrid Framework for Corporate Payment-Anomaly Detection and Working-Capital Decision Support

Tanisi Das

tanisidas.21@gmail.com

Independent Researcher

Research Article

Keywords: explainable artificial intelligence, financial anomaly detection, corporate payments, Isolation Forest, Bayesian logistic regression, financial risk, working-capital management, human-in-the-loop decision support

Posted Date: October 22nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11144714/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

PayGuard AI: An Explainable Hybrid Framework for Corporate Payment-Anomaly Detection and Working-Capital Decision Support

Author: Tanisi Das

Independent Researcher, Dubai, United Arab Emirates

Correspondence: tanisidas.21@gmail.com

Abstract

Corporate Accounts Payable processes involve high transaction volumes, diverse suppliers, and complex payment conditions, creating challenges for manual fraud screening. Existing detection methods commonly prioritise classification performance while providing limited interpretability, uncertainty quantification, or financial decision support. This study introduces PayGuard AI, an explainable hybrid framework integrating business rules, Isolation Forest anomaly detection, logistic regression, Bayesian uncertainty estimation, bootstrap review-exposure analysis, and Monte Carlo working-capital simulation. The framework was evaluated on 5,600 synthetic supplier-payment transactions containing a 5% controlled anomaly rate across seven categories. On held-out data, the hybrid achieved an area under the receiver operating characteristic curve (ROC-AUC) of 0.982 versus 0.989 for rules alone, but higher average precision (0.770 versus 0.680); its precision, recall, and specificity were 0.848, 0.667, and 0.994, respectively. Sensitivity varied by anomaly type, with weak detection of exact and near duplicates. The statistical models provided interpretable feature associations and approximate posterior uncertainty estimates. Bootstrap resampling produced a mean aggregate review exposure of approximately 302,564 synthetic monetary units (SMU), while Monte Carlo analysis illustrated uncertainty across Days Payable Outstanding scenarios. PayGuard AI offers a proof-of-concept approach that combines transaction screening, interpretability, uncertainty representation, and financial decision support. Real-world validation and organisational calibration are still needed before operational deployment.

Keywords: explainable artificial intelligence; financial anomaly detection; corporate payments; Isolation Forest; Bayesian logistic regression; financial risk; working-capital management; human-in-the-loop decision support

1. Introduction

Corporate Accounts Payable (AP) processes involve reviewing supplier invoices and arranging payments for purchased goods and services. These processes can be exposed to payment fraud and processing errors. In the 2025 Payments Fraud and Control Survey by the Association for Financial Professionals (AFP), 79% of respondents reported that their organisations had experienced attempted or actual payments fraud during 2024 . Payment fraud can result in direct financial losses, operational disruptions, audit failures, reputational harm, and legal consequences. In the 2024 Report to the Nations, surveyed fraud examiners estimated that organisations lose 5% of annual revenue to occupational fraud. The median loss among the occupational-fraud cases examined in the report was US\$145,000 . The complexity of processing large volumes of invoices is further compounded by factors such as diverse suppliers, multiple currencies, and varying payment terms, making comprehensive manual review of each payment challenging. Consequently, there is growing interest in automated fraud detection systems.

The Accounts Payable workflow consists of several stages: purchase order (PO) creation, receipt of goods, supplier invoice submission, invoice approval, and payment. Fraudulent activities can occur at any of these stages. Duplicate invoices may arise from processing errors or deliberate attempts to obtain repeated payment; identifying a duplicate does not, by itself, establish fraudulent intent . Another prevalent type is supplier impersonation, where fraudsters pose as legitimate suppliers to request payments. This often manifests as Business Email Compromise (BEC) , in which perpetrators gain access to or impersonate a supplier's email account. These forms of fraud frequently mimic legitimate supplier activities, making detection by finance teams challenging through manual review alone. Consequently, organisations implement internal controls and use automated fraud detection systems to help professionals identify fraudulent transactions.

Corporations have traditionally used internal business controls to detect fraudulent activity . A rule-based system evaluates whether a transaction violates predefined business rules designed to identify suspicious or potentially fraudulent behaviour . The operation of such a system involves checking whether a condition is violated; if so, the transaction is flagged for manual review. This indicates a deterministic process in which the same payment will always yield the same outcome. For example, duplicate invoice matching may be detected by comparing invoice details, supplier information, and invoice amounts with previous payments. Paying an identical invoice would flag the transaction. Similarly, regarding purchase order matching, an invoice may be compared with the approved purchase order. If the invoice amount exceeds the purchase order amount, the transaction is flagged. The use of rule-based systems is popular due to their high interpretability and transparency, with every flagged transaction providing a clear explanation for its alert based on explicit business rules . This allows for easier and quicker auditing without complex computation. However, rule-based systems pose

several limitations. Since their working depends on explicitly defined business rules, only anticipated fraud may be detected, without accounting for any new techniques that fraudsters may develop. As such, rules would require continuous updating, with each method of fraud requiring its own fraud pattern. Moreover, certain transactions may result in false positives. Corporations often have payment thresholds as a business rule. However, situations such as urgent system maintenance may legitimately exceed payment thresholds and still be flagged. Rules limited to individual-payment attributes may overlook patterns across related transactions unless historical or relational checks are explicitly included. To overcome the limitations of such a deterministic system, machine learning has been increasingly incorporated into fraud detection.

In contrast to a rule-based system, machine learning identifies patterns from data rather than relying on predetermined business rules. Supervised learning is a form of machine learning, which involves evaluating historical transactions and training a model to learn the relationship between these transactions and whether they are flagged as fraudulent. Common algorithms used for such evaluation include logistic regression, decision trees, random forests, XGBoost, and neural networks. This allows the model to predict whether future transactions may be fraudulent. Depending on the algorithm employed, supervised models can learn complex relationships and identify informative predictors, while their performance generally benefits from sufficiently large and representative labelled datasets. However, its setup may require a large amount of correctly labelled fraud examples, which can be scarce.

Furthermore, fraudulent behaviour may change over time, so historical fraud cases may inaccurately represent future cases. Corporations do not often have a significant amount of confirmed fraud cases. Therefore, the discussion shifts to unsupervised learning, which allows for anomaly detection. The model for such a process learns normal behaviour, observes transactions that deviate significantly from it, and flags them as anomalies. Isolation Forest is an example of such a model. Anomaly detection can identify unusual transaction patterns that are not explicitly represented in predefined rules, although adaptation to changing behaviour requires appropriate updating or retraining. Moreover, it proves useful in imbalanced datasets. However, not every detected anomaly is fraudulent, and legitimate payments may be unusual - indicating the importance and necessity of human investigation. Research suggests combining traditional business controls with machine learning models to leverage the interpretability and transparency of rule-based systems alongside the adaptability of data-driven models. Research on interpretable machine learning emphasises the need to explain and evaluate model behaviour [9, 10]. Building on this motivation, the present study combines payment-anomaly screening with uncertainty summaries and illustrative financial decision support within a single proof-of-concept workflow. In operational contexts, finance professionals require not only effective anomaly identification, but also interpretable predictions and reliable confidence estimates to facilitate informed payment decisions. These unresolved challenges provide the rationale for the framework introduced in this study.

To address the limitations of current fraud detection systems, this research introduces PayGuard AI, an explainable hybrid decision-support framework for corporate payment fraud detection. PayGuard AI emphasises both effective payment-anomaly identification and transparent interpretability of results, quantifies predictive uncertainty, and delivers actionable financial insights to inform operational decision-making. The proposed framework consists of several components, each designed to address a specific limitation of existing fraud detection approaches. Business rules and Isolation Forest provide complementary numerical inputs to the hybrid score: the rules encode predefined payment controls, while Isolation Forest measures statistical anomalousness. Logistic regression is applied to assess the influence of transaction characteristics on the likelihood of planted anomaly status. Bayesian logistic regression is used to estimate predictive uncertainty and provide probabilistic interpretations. Bootstrap resampling quantifies uncertainty in aggregate review exposure, while Monte Carlo simulation evaluates illustrative treasury and working-capital scenarios. Together, these stages combine business controls, machine learning, probabilistic modelling, and financial simulation within a unified framework that extends beyond basic fraud detection and classification. The framework is evaluated using a synthetic corporate payment dataset designed to simulate realistic supplier transaction scenarios. Experimental results demonstrate the framework's ability to prioritise suspicious payments under controlled synthetic conditions while providing interpretable risk assessments, uncertainty estimates, and illustrative financial decision-support outputs.

Guided by these objectives, this study investigates the following research questions:

RQ1: How effectively can a hybrid framework combining rule-based controls and machine learning identify anomalous corporate payment transactions under controlled synthetic conditions?

RQ2: Can interpretable statistical modelling and Bayesian inference improve the transparency of fraud-detection results by providing interpretable feature relationships and explicit uncertainty estimates?

RQ3: Can integrating financial risk analysis through Monte Carlo simulation provide decision-relevant insights for treasury and working-capital decision-making beyond conventional fraud classification?

The structure of this paper is as follows: Section 2 reviews related literature, Section 3 outlines the proposed methodology, Section 4 details the experimental evaluation, Section 5 discusses the findings, and Section 6 provides the conclusion.

2. Literature Review

Research on financial fraud detection systems has progressed from deterministic, rule-based approaches to data-driven machine learning models that can identify emerging fraud patterns. In recent years, there has been increased focus on explainable artificial intelligence (XAI), probabilistic modelling, and decision-support systems as advanced

techniques for fraud detection. These methods enhance transparency and interpretability, thereby increasing the practical applicability of fraud detection systems. The subsequent sections examine traditional fraud detection methods, recent developments in machine learning, explainable artificial intelligence, and financial decision-support techniques, which together provide the rationale for the proposed framework.

2.1. Rule-Based Fraud Detection

Rule-based systems were among the earliest methods employed for financial fraud detection . These systems initially relied on expert knowledge rather than statistical learning. Consequently, finance professionals developed deterministic business rules to support internal control procedures. Such rules identify fraudulent behaviour through indicators including duplicate invoices, unusual payment amounts, unauthorised suppliers, purchase order mismatches, and abnormal payment timing. Rule-based systems continue to be widely used because they are consistent with established business practices . Recent research in fraud detection increasingly considers rule-based controls as a single component within comprehensive systems that also utilise data-driven analytical techniques .

Rule-based systems remain relevant due to several distinct advantages. They offer highly interpretable and transparent decision-making, as each flagged transaction corresponds to an explicitly defined business rule . The deterministic structure of these systems produces consistent outcomes when identical rules are applied to transactions with the same characteristics . The transparency of these systems facilitates auditing, enabling investigators to readily identify the rationale behind flagged payments. Additionally, such systems support financial governance and auditing requirements by providing traceable and well-defined decision criteria . Furthermore, rule-based systems are straightforward to implement and require less historical data compared to machine learning-based approaches .

Although rule-based systems are foundational in fraud detection, they exhibit several significant limitations. Their deterministic design restricts detection to fraud scenarios that experts have previously anticipated. These systems lack adaptability to emerging fraud techniques, requiring manual updates to business rules for new threats . With the expansion of fraud scenarios, the maintenance and updating of a growing set of manually specified rules presents significant operational challenges. Additionally, rule-based systems may generate false positives when legitimate transactions breach predefined thresholds, leading to increased investigation costs . Furthermore, assessment rules that focus solely on individual payments, without considering historical or relational behaviour, often overlook subtle patterns that emerge across multiple transactions or entities. These challenges have motivated researchers to investigate data-driven fraud detection methods that are adaptable and informed by historical payment data .

2.2. Machine Learning-Based Fraud Detection

Financial fraud detection increasingly incorporates machine learning and data mining because these approaches can efficiently analyse large transaction datasets and identify statistical relationships within historical transaction data. They have been applied across various fraud domains, including payment-card, insurance, financial-statement, banking and other transaction-based fraud. For example, Ravisankar et al. investigated data-mining techniques for financial-statement fraud detection and feature selection. Ngai et al. identified logistic models, neural networks, Bayesian belief networks and decision trees among the principal techniques examined in the literature. Phua et al. additionally highlighted challenges associated with automated data-driven fraud detection, including class imbalance, evolving fraud strategies, limited labelled data and the financial consequences of incorrect predictions. Machine-learning approaches to fraud detection can be broadly categorised as supervised or unsupervised. Recent reviews indicate that supervised methods remain predominant in financial fraud research, although unsupervised methods have received increasing attention where reliable fraud labels are limited. Leevy et al. compared one-class and binary classifiers on imbalanced credit-card and Medicare fraud datasets using receiver operating characteristic and precision–recall measures, finding that binary classification outperformed one-class classification in their experiments.

The effectiveness of supervised learning methods is influenced by the quantity, quality and representativeness of labelled examples. Legitimate transactions generally outnumber fraudulent ones, producing highly imbalanced datasets. Such imbalance may allow models to demonstrate high overall accuracy while performing poorly on the minority fraud class. Consequently, fraud-detection studies commonly report class-sensitive measures such as precision, recall, F1-score and area under the receiver operating characteristic curve alongside overall accuracy. Additionally, supervised models may lose effectiveness when emerging fraud patterns differ from the historical cases used during training. Logistic regression is an established statistical method in financial fraud detection. It estimates the probability of a binary outcome by modelling the relationship between predictor variables and the log-odds of that outcome. Its principal advantage is interpretability, as coefficient signs and magnitudes indicate the direction and strength of associations between individual predictors and fraud likelihood. Logistic regression therefore provides a useful baseline in settings where explanatory insight is required alongside classification. However, the standard model assumes a specified functional relationship between predictors and the log-odds of fraud, limiting its ability to represent complex nonlinear interactions unless transformations or interaction terms are explicitly introduced.

Random Forest constructs an ensemble of decision trees using randomly sampled observations and predictor variables. This method effectively models nonlinear relationships and has been widely applied in financial fraud detection. However, Random Forest provides lower interpretability compared to logistic regression, as

individual predictions are difficult to explain . Additionally, its performance may be compromised by severe class imbalance unless appropriate mitigation strategies are implemented . XGBoost constructs decision trees sequentially to minimize prediction errors and incorporates regularisation to enhance generalisation . This algorithm has demonstrated robust performance on structured transaction data. However, its predictive behaviour is less interpretable than that of logistic regression and is highly dependent on hyperparameter selection and the quality of training data . Neural networks learn complex nonlinear relationships and have demonstrated strong predictive performance in financial fraud detection. Their application has also extended to audit-oriented decision support, with Bhattacharya et al. proposing an artificial neural network (ANN) system incorporating Benford's Law to support the identification of potentially irregular financial data. Their flexibility enables them to model complicated transaction patterns but requires substantial data and computational resources . Furthermore, their limited interpretability reduces their suitability for audit-sensitive financial applications .

Corporations commonly process large volumes of transaction data while possessing relatively few reliable fraud labels, motivating the use of unsupervised learning approaches . Unlike supervised classification, unsupervised anomaly detection identifies observations or patterns that deviate substantially from expected behaviour without requiring complete fraud labels. Such anomalies may indicate fraud, but legitimate transactions can also deviate from normal behaviour. Anomaly detection can therefore serve as a preliminary screening and prioritisation mechanism rather than definitive evidence of fraud, with flagged transactions requiring further investigation . Isolation Forest is an unsupervised learning model that detects anomalies by repeatedly partitioning the feature space using randomly selected variables and split values. Anomalies are generally isolated with fewer partitions, while regular values are found within dense, normal regions. Unlike many alternative methods, Isolation Forest does not require an initial estimation of the complete normal data distribution, which increases computational efficiency for large datasets. Additionally, it is well-suited for screening purposes, as it assigns an anomaly score to each payment without the need for predefined fraud labels. However, this approach identifies only statistically unusual cases rather than confirmed instances of fraud. Its practical performance nevertheless depends on the relevance of the selected features, the representativeness of the training data and the threshold used to convert anomaly scores into alerts .

The Local Outlier Factor (LOF) is an unsupervised learning model that detects anomalies by comparing the local densities of observations. An observation is classified as anomalous if it resides in a region with significantly lower density than its neighbours. This approach is particularly suitable when normal transaction behaviour forms distinct groups, such as suppliers exhibiting different patterns. However, its performance depends on the size and distance metrics used to define the neighbourhood. Additionally, LOF can be computationally intensive for large datasets due to the need to calculate neighbourhood relationships . Autoencoders are neural networks designed to

reconstruct input data through a compressed internal representation. For anomaly detection, these models are typically trained primarily on normal transactions. A transaction is considered anomalous if it yields a large reconstruction error, indicating that the model cannot represent it as effectively as typical behaviour. Autoencoders capture nonlinear relationships and are particularly effective when transaction datasets contain complex feature interactions. The advantages of these neural networks have led recent financial fraud literature to examine autoencoders and other deep-learning anomaly detection techniques. However, autoencoders require greater computational resources and more extensive tuning than alternative models. Additionally, they offer limited interpretability, as they do not provide clear business explanations. Similar to the Isolation Forest and Local Outlier Factor (LOF) models, autoencoders identify unusual observations rather than directly confirming instances of fraud .

Unsupervised learning methods operate effectively without heavy reliance on labelled fraud cases. These methods facilitate the identification of previously unseen behaviours; however, such detections serve as indicators of anomalous behaviour rather than definitive classifications of fraud. The potential for false positives necessitates the integration of anomaly scores with business controls, interpretable evidence, and human review.

Recent studies indicate that researchers increasingly combine rule-based and machine-learning methods rather than relying on a single algorithm, as these approaches offer complementary strengths . Hybrid approaches integrate two or more of these methods, ensuring that no single mechanism determines the complete outcome. Typically, business rules are employed for initial screening, with machine learning applied in subsequent stages for further prioritisation. This strategy enables broader coverage, allowing for the simultaneous investigation of known policy violations and statistical anomalies . Incorporating domain knowledge further ensures that model results are interpreted in conjunction with accounting controls, thereby enhancing the relevance of data-driven systems . Despite the advantages of combining multiple methods, hybrid systems do not always ensure explainability. Such approaches may generate risk scores and classifications without clarifying the level of confidence in the results. Current research often evaluates detection performance using classification metrics , yet these do not assess the financial loss associated with suspicious payments. Furthermore, these metrics fail to provide sufficient guidance on how flagged or anomalous payments should influence cash-flow planning, payment timing, or working-capital decisions. Although hybrid systems enhance fraud detection performance, many still depend on complex predictive models with internal decision-making processes that are challenging for finance professionals to interpret. These limitations motivate the development of PayGuard AI. This system combines rule-based screening, machine learning, interpretable statistical modelling, uncertainty estimation, and financial simulation in a unified decision-support framework.

2.3. Explainability in Financial AI

The increasing complexity of machine learning models has led to the prevalence of black-box systems, which generate predictions without offering explanations for their decision-making processes. Deep neural networks and ensemble tree models exemplify such black-box approaches. While these systems often achieve high predictive performance, they lack interpretability, a critical requirement in financial fraud detection. In contrast to other machine learning applications, financial decisions carry legal, regulatory, and operational implications. Therefore, prediction accuracy alone is inadequate. Černevičienė and Kabašinskas reviewed 138 studies on explainable artificial intelligence in finance, documenting differences in explanation methods across financial applications and identifying unresolved challenges in their development and use .

The objective of explainability extends beyond merely describing model functionality; it enables users to interpret and trust predictions for informed decision-making. Explanations help users assess whether model outputs are reasonable and consistent with established domain knowledge. Additionally, they facilitate the identification of cases where models may depend on misleading or unintended data relationships. Interpretable predictions enhance transparency and support accountability during the investigation of suspicious transactions. Numerous researchers advocate for the use of interpretable models over black-box models when comparable predictive performance is attainable. Rudin further contends that in high-stakes domains such as finance and healthcare, interpretable models should be prioritised over post-hoc explanations of black-box systems. Consequently, explainability not only enhances user comprehension but also fosters organisational trust and responsible deployment of artificial intelligence . Sabharwal et al. applied explainable artificial intelligence, including SHAP-based explanations, to financial anomaly detection, illustrating how model outputs can be accompanied by information intended to support managerial interpretation.

While machine learning models can provide explanations for their predictions, they often fail to convey the level of confidence associated with those predictions . For instance, transactions assigned identical fraud probabilities may exhibit varying degrees of predictive uncertainty. Predictive uncertainty is crucial in fraud detection, as false alerts can delay legitimate payments and missed alerts may result in financial losses. Bayesian statistical methods offer a probabilistic framework that quantifies uncertainty by estimating probability distributions over model parameters, rather than relying on single fixed values . These models assess the degree of confidence in each prediction, enabling finance professionals to distinguish between predictions supported by robust evidence and those warranting further scrutiny . Despite increasing attention to explainable artificial intelligence, limited research in fraud detection integrates explainability with uncertainty estimation in operational decision support systems. These gaps underscore the need to incorporate explainable statistical modelling and Bayesian uncertainty estimation within PayGuard AI, thereby enabling finance professionals to

interpret predictions and assess their associated confidence. In addition to interpretability and predictive confidence, it is essential for finance professionals to assess the financial consequences associated with suspicious transactions.

2.4. Financial Risk Analysis and Decision Support

Current research in fraud detection evaluates models using predictive performance measures that distinguish between fraudulent and legitimate transactions. Predictive performance measures alone do not establish the operational or financial consequences of acting on suspicious transactions. Following fraud detection, financial professionals must decide whether specific payments should be delayed, investigated, approved, or rejected. This decision-making requires balancing fraud risk with business continuity, supplier relationships, liquidity, and operational efficiency. Therefore, fraud detection should be regarded as one stage within a broader financial decision-making process rather than as a single event .

In financial decision-making, institutions quantify uncertainty. Financial risk management enables the evaluation of multiple possible scenarios with their associated probabilities, rather than relying on a single predicted outcome. This approach supports decision-making under uncertainty by assessing the potential financial consequences of alternative outcomes. Instead of relying exclusively on deterministic estimates, risk analysis allows organisations to assess a spectrum of potential financial outcomes and the probability associated with each. This approach offers decision-makers a more thorough understanding of potential losses and supports risk-informed payment and treasury management decisions. By quantifying uncertainty, financial professionals can determine appropriate responses. Integrating financial analysis can provide additional decision context, but its practical value requires evaluation against organisational outcomes .

Monte Carlo simulation evaluates possible future outcomes by repeatedly sampling random values from probability distributions. The Monte Carlo simulation generates a distribution of outcomes rather than a single loss estimate, enabling analysts to observe both expected losses and the variability associated with uncertain events.

Consequently, Monte Carlo methods are widely used in financial engineering and quantitative risk management . Value at Risk (VaR) is a quantile of the loss distribution for a specified time horizon and confidence level . Similarly, Conditional Value at Risk (CVaR) estimates the average loss beyond the VaR threshold, providing additional insight into extreme financial outcomes and offering a more informative measure than VaR alone . These risk measures demonstrate how distributional analysis can characterise uncertainty and tail outcomes; however, their interpretation requires calibrated loss distributions, which may not be available in proof-of-concept fraud-detection settings.

Within the Accounts Payable process, fraud detection decisions have broader implications than prevention alone. Delays in legitimate payments can affect supplier

relationships, payment discounts, and working capital performance, while accepting fraudulent transactions can result in financial losses. Therefore, finance professionals must balance fraud risk with liquidity and operational management. Research on working-capital management has reported associations between payment timing and firm profitability . Treasury departments must evaluate the trade-off between prompt payment release, which supports supplier relationships, and payment delays necessary for investigating potential fraud. Addressing these competing objectives requires quantitative evidence concerning both fraud risk and the financial consequences associated with various payment decisions. Despite this, limited research has integrated predictive uncertainty, financial simulation, and treasury decision support into a unified framework. This gap motivates the financial risk analysis component of PayGuard AI.

3. Materials and Methods

3.1. Research Design

This study employs a quantitative experimental methodology to design and evaluate an explainable hybrid fraud-detection framework for corporate Accounts Payable transactions. The proposed framework integrates rule-based controls, unsupervised anomaly detection using the Isolation Forest model, interpretable statistical modelling through logistic regression, Bayesian uncertainty estimation, and financial risk simulation via Monte Carlo simulation and review-exposure analysis. Each component addresses specific limitations identified in the existing literature. Rule-based controls enhance transparency and alignment with business controls by screening for known fraud patterns, addressing the need for clear explanations. Isolation Forest contributes a measure of statistical unusualness to the hybrid ranking. At the selected weighting and alert threshold, however, statistical unusualness alone cannot trigger an alert; sufficient rule-based evidence is also required. Logistic regression evaluates the relationship between transaction characteristics and anomaly likelihood, providing an interpretable assessment of the factors associated with suspicious transactions. Bayesian logistic regression estimates posterior distributions for model parameters and predictions, enabling predictive uncertainty to be quantified and addressing the lack of explicit uncertainty estimation in standard fraud classification. Financial simulation translates analytical outputs into financial consequences, supporting decision-making under uncertainty and extending research beyond mere classification to consider subsequent financial decisions.

The framework is evaluated using a synthetic dataset for corporate payments, as real data containing confirmed fraud cases are seldom publicly available due to privacy and security constraints . Jensen et al. provide a related example of synthetic financial data for benchmarking anti-money-laundering methods, although their dataset addresses a different task from corporate supplier-payment screening. The synthetic dataset was constructed to emulate realistic supplier transaction behaviour, with controlled insertion of predefined anomaly scenarios. Both dataset generation and experimental procedures were implemented in Python.

The analyses were implemented in Python using NumPy, pandas, scikit-learn, PyMC, ArviZ and Matplotlib. The supplied Bayesian posterior file records PyMC 5.26.1 and ArviZ 0.23.4. Software versions recorded for the final export environment are provided in analysis/environment_versions.txt within Supplementary File S1. The supplementary archive also includes the synthetic dataset, source code, numerical outputs and figure-generation materials.

The Excel evaluation workbook uses 20,000 resamples for aggregate review exposure and 20,000 draws for the selected working-capital scenario. At its default settings, its resampling method and input distributions match the corresponding Python analyses. Excel generates independent random draws on recalculation, so individual numerical summaries may differ because of Monte Carlo variation. The reported manuscript results are taken from the saved Python run with the specified random seeds.

3.2. Overall Framework Architecture

Fig. 1 illustrates the architecture of PayGuard AI, a multi-stage analytical and decision-support framework. The process begins with a detection layer that processes corporate supplier-payment records. Each raw payment item is subjected to feature engineering to derive numerical indicators for transaction amount, payment behaviour, and control-related characteristics. The engineered data is then evaluated using two complementary detection mechanisms. First, deterministic business-rule screening identifies transactions exhibiting predefined fraud indicators. Subsequently, the Isolation Forest model assesses whether each transaction statistically deviates from patterns observed in the training data. The outputs of these mechanisms are combined to generate a hybrid risk score, which determines the prioritisation of transaction investigations. Each flagged transaction is accompanied by a human-readable alert explanation and review-exposure information to facilitate the investigation.

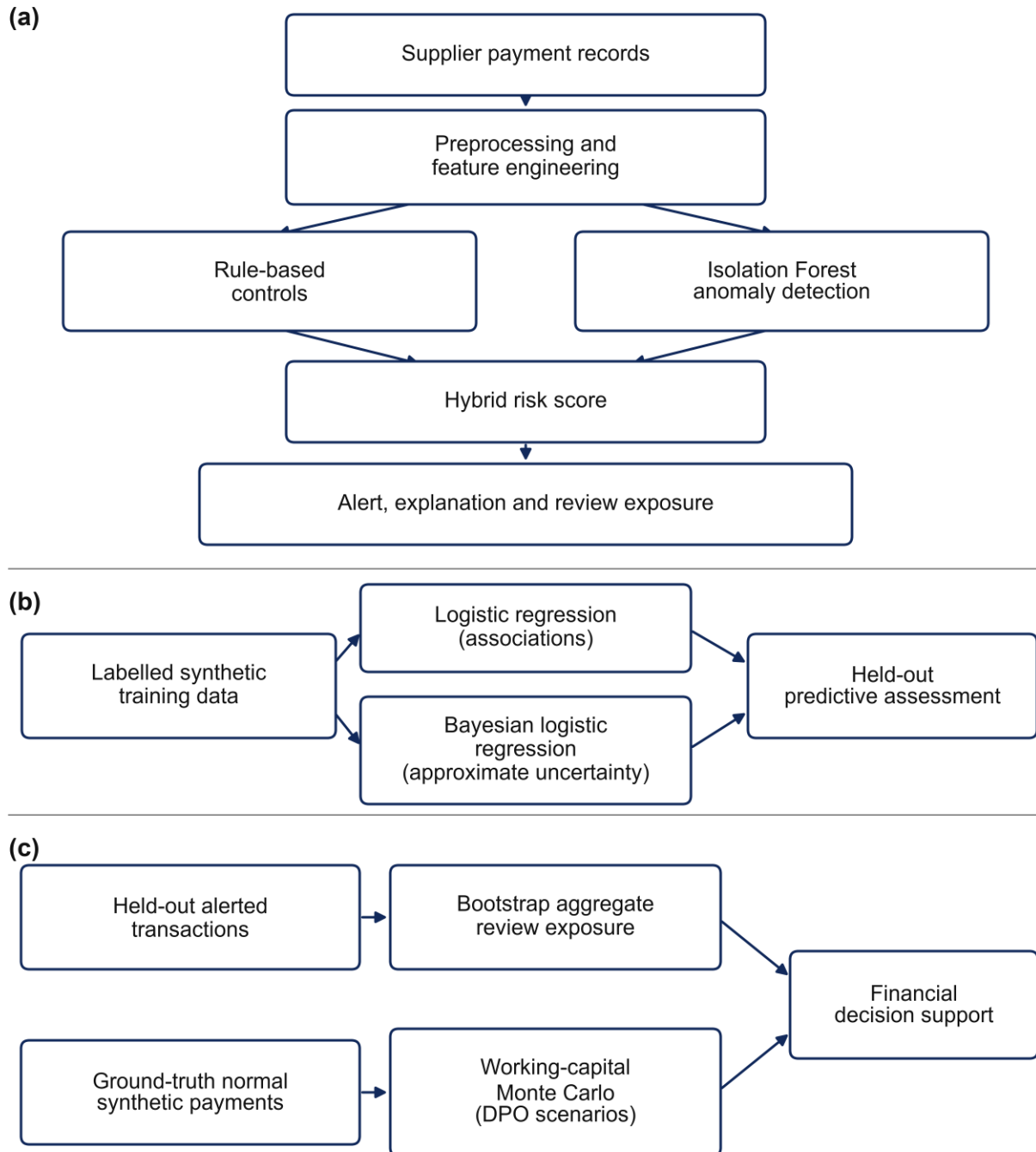


Fig. 1 *Architecture of PayGuard AI: (a) operational detection and alert generation; (b) separately fitted statistical models and held-out assessment; (c) bootstrap review-exposure analysis and Monte Carlo working-capital simulation*

Beyond the detection layer, the framework incorporates interpretable statistical modelling to evaluate anomalous transaction behaviour and associated relationships. Logistic regression is applied to the previously generated synthetic training data to

assess the strength of association between each payment feature and the likelihood of a planted anomaly. Coefficients and odds ratios provide interpretable measures of feature influence. Bayesian logistic regression extends this approach by estimating posterior distributions for model parameters and predictive probabilities, thereby quantifying uncertainty in estimated feature relationships and individual transaction predictions. These statistical models serve as complements to, rather than replacements for, the rule-based and Isolation Forest detection mechanisms.

The final layer of the framework supports financial decision-making by evaluating the potential consequences of flagged payments and payment-timing strategies. For flagged transactions, review exposure is calculated from directly measurable duplicate or overpayment amounts. For cases without directly observable exposure, an illustrative scenario estimate is provided. Bootstrap simulation is used to assess the distribution of aggregate review exposure among held-out flagged transactions. Separately, Monte Carlo simulation evaluates the potential working-capital implications of extending Days Payable Outstanding by 10, 20, and 30 days, considering uncertainty in cost of capital, early-payment discounts, and supplier-friction costs. These analyses extend PayGuard AI beyond transaction classification by incorporating potential financial implications of payment risk and treasury decisions. The simulation outcomes are intended as proof-of-concept decision-support analyses rather than calibrated forecasts for a specific organisation.

3.3. Synthetic Dataset Generation

The synthetic dataset employed in this framework was programmatically generated to simulate an Accounts Payable environment. A fixed random seed of 42 was used for synthetic data generation. Each row represents an individual supplier payment transaction and contains information relevant to the primary stages of payment screening, including supplier details, purchase order data, invoice information, payment specifics, currency, payment terms, and supplier bank account details. The dataset primarily comprises normal transactions, with a controlled number of anomalies deliberately inserted. These planted anomalies provide known ground-truth labels, enabling evaluation of whether the PayGuard AI detection framework accurately identifies anomalous transactions. The labels, indicated by the `is_anomaly` flag in the code, denote synthetic anomaly status and do not imply that a transaction has been confirmed as fraudulent. All transactions are synthetic. The raw payment, invoice and purchase-order fields contain nominal amounts associated with synthetic currency codes. For monetary aggregation, amounts are converted to a common synthetic base unit, reported as synthetic monetary units (SMU), using fixed illustrative factors of 1.0000 for USD, 0.2723 for AED, 1.0900 for EUR and 1.2700 for GBP. These factors are modelling assumptions rather than observed transaction-date exchange rates. Payment and invoice amounts are converted using their respective currency codes before review exposure is calculated. Annualised spending and working-capital outcomes use the

same base unit. The legacy code names `amount_usd` and `FX_TO_USD` are retained for compatibility and refer to this synthetic conversion convention.

Suppliers in the dataset are distributed across six procurement categories: aircraft components, maintenance, information technology, ground services, professional services, and facilities. This diversity introduces variation in transaction data, avoiding uniformity among payments. Additional realism is achieved through variation in transaction monetary values, logical relationships between purchase orders and invoices, alignment of paid amounts with invoice amounts, consistent associations between suppliers, currencies, and bank accounts, and payment dates that correspond to invoice dates and payment terms. Consequently, payment behaviour varies naturally among legitimate transactions. These features collectively establish a heterogeneous baseline of standard supplier-payment behaviour, facilitating the investigation of anomalous transactions.

The data generator produces several categories of controlled anomalies. For duplicate invoice anomalies, characteristics of a normal transaction, including supplier, purchase order, invoice, currency, payment, and bank account information, are replicated to create transactions resembling the original. Two variants are generated: an exact duplicate, which retains the original invoice number, and a near duplicate, in which a minor modification is introduced to the invoice number while preserving high similarity to the original. Overpayment anomalies are created by increasing the payment amount relative to the corresponding invoice, resulting in an abnormal paid-to-invoice amount ratio. This tests the framework's ability to detect overpayments that exceed invoiced obligations. Invoice-to-purchase-order mismatch anomalies are introduced by deliberately increasing the invoice amount relative to the purchase order, producing an abnormal invoice-to-purchase-order ratio and representing a control violation. Recent supplier bank-account change anomalies are generated by assigning a new bank-account identifier and activating the recent bank-change indicator. Currency mismatch anomalies are generated by altering the payment currency so that it is inconsistent with the invoice currency. Rapid payment anomalies are created by setting the payment date to zero or one day after the invoice date, representing payment substantially earlier than standard payment terms.

Controlled anomaly injection is a key feature of the dataset. Anomaly locations are deliberately selected so that each selected transaction is assigned a single planted anomaly category. Duplicate transactions are generated from otherwise normal source transactions, allowing observed model behaviour to be more directly associated with the intended anomaly scenario. Each transaction is assigned a ground-truth label indicating whether it is normal or a planted anomaly. These labels are used for supervised statistical analysis and quantitative evaluation but are not required for unsupervised Isolation Forest model fitting. Because the dataset is entirely synthetic, performance measures reflect the methodology's effectiveness under controlled experimental

conditions and do not indicate expected fraud-detection performance in real corporate environments.

3.4. Feature Engineering and Preprocessing

Feature engineering constitutes a critical element of the detection framework. It is applied to raw transaction records, including invoice and payment dates, payment terms, currencies, and bank-change information, which may provide limited information when considered independently. Through feature engineering, these raw values are transformed into derived features that capture relationships among the transaction attributes. The primary objective is to enable explicit representation of these relationships. For instance, the absolute value of a payment alone does not indicate overpayment; it must be compared with the corresponding invoice amount to yield a meaningful metric. Such derived values provide numerical representations that support rule-based screening, anomaly detection, and statistical analysis within the framework. Table 1 summarises the engineered transaction features, their calculation and their purpose.

Feature	Calculation/Representation	Purpose
Paid-to-invoice ratio	$\text{Paid Amount} / \text{Invoice Amount}$	Measures whether the payment exceeds the invoiced obligation
Invoice-to-PO ratio	$\text{Invoice Amount} / \text{PO Amount}$	Measures whether invoiced expenditure exceeds the authorised purchase amount
Days to pay	$\text{Payment Date} - \text{Invoice Date}$	Represents the elapsed time between invoicing and payment
Payment-term deviation	$\text{Days to Pay} - \text{Payment Terms}$	Measures how early or late payment occurs relative to agreed terms
Log payment amount	$\ln(1 + \text{Paid Amount in SMU})$	Reduces the influence of highly skewed monetary values
Weekend-payment indicator	Binary indicator derived from payment date	Identifies payments occurring on Saturday or Sunday
Recent bank-change indicator	Binary indicator	Represents whether supplier bank details were recently changed
Currency-mismatch indicator	Binary indicator	Represents disagreement between invoice and payment currencies

Table 1. Engineered transaction features

The paid-to-invoice and invoice-to-purchase-order ratios are calculated from the nominal amount fields used by the synthetic generator. For amounts in the same currency, a ratio above one indicates that the numerator exceeds the denominator. For invoice/payment currency mismatches, the nominal paid-to-invoice ratio is retained only as a synthetic control feature and is not interpreted as an exchange-rate-adjusted overpayment measure; the currency-mismatch indicator captures the currency-code disagreement separately. Review exposure is valued independently using amounts converted to the common SMU base, as described in Section 3.7.1.

Payment timing is characterised by both days-to-pay and payment-term deviation. Days-to-pay quantifies the interval between invoice and payment dates, while payment-term deviation compares this interval to the supplier's agreed payment terms. A negative payment-term deviation indicates payment occurred before the agreed terms, whereas a positive value indicates payment after the expected date. Incorporating both features enables the framework to distinguish between the absolute speed of payment and deviations from supplier-specific terms.

The payment amount is represented using $\ln(1 + P_i)$, where P_i is the payment amount converted to the common SMU base using its payment-currency code. This transformation mitigates the influence of exceptionally large payments while preserving distinctions in transaction magnitude. It is particularly advantageous when payment values span a wide range, as the transformation compresses extreme monetary values and reduces skewness in the resulting feature distribution.

The weekend-payment indicator is distinct from the planted anomaly categories discussed in Section 3.3. Unlike ground-truth anomalies, weekend payments were not deliberately introduced but arose from the payment dates generated during normal dataset construction and were preserved as a contextual behavioural indicator.

Feature scaling was performed according to the requirements of each modelling component. For Isolation Forest anomaly detection, continuous features were transformed using robust scaling based on the training-set median and interquartile range, reducing the influence of extreme transaction values. For the logistic and Bayesian logistic regression analyses, continuous predictors were standardised using $z = (x - \mu)/\sigma$, where μ and σ denote the corresponding training-set mean and standard deviation. Binary indicators retained their original 0/1 representation in both cases. Rule-based screening operated on the original interpretable feature values to ensure that predefined business thresholds retained their intended meaning. All scaling parameters were estimated exclusively from the training data and subsequently applied to held-out observations, thereby preventing preprocessing-related data leakage.

3.5. Hybrid Fraud Detection Framework

After feature engineering, the PayGuard AI framework analyses each transaction using both deterministic business rules and unsupervised anomaly detection. These mechanisms are combined because they offer complementary strengths: business rules

identify transactions exhibiting known, explicitly defined control violations, whereas the Isolation Forest model detects transactions that statistically deviate from normal behaviour in the training data without the need for anomaly labels. The outputs from both methods are integrated into a single hybrid risk score, which is subsequently used to prioritise transactions for further investigation. This approach maintains transparency for traditional controls while leveraging the adaptability of data-driven anomaly detection.

3.5.1. Rule-Based Screening

The rule engine assesses each transaction using a predefined set of payment-control conditions. Each rule corresponds to a specific, interpretable transaction characteristic and is assigned a weight that indicates its relative severity. If a transaction triggers a rule, it contributes a predetermined number of risk points according to the assigned weight. The total risk points from all triggered rules constitute the rule-based component of the hybrid risk assessment. Table 2 summarises the screening conditions and their assigned risk contributions.

Rule	Condition	Interpretation	Risk Contribution
Potential duplicate invoice	Exact duplicate or near-duplicate invoice identified	Indicates a potentially repeated supplier payment	35
Payment exceeds invoice	$\text{Paid Amount} > \text{Invoice Amount} \times 1.01$	Nominal paid amount exceeds nominal invoice amount by more than 1%	35
Invoice exceeds purchase order	$\text{Invoice Amount} > \text{PO Amount} \times 1.05$	Nominal invoice amount exceeds nominal purchase-order amount by more than 5%	28
Recent bank-account change	$\text{bank_changed_recently} = 1$	Supplier bank-account details were recently changed	32
Invoice/payment currency mismatch	$\text{Invoice Currency} \neq \text{Payment Currency}$	Payment currency differs from the corresponding invoice currency	30
Unusually rapid payment	$\text{Days to Pay} \leq 1$	Payment occurs no more than one	18

Rule	Condition	Interpretation	Risk Contribution
		day after the invoice date	
Weekend payment	Payment Date = Saturday or Sunday	Payment occurs during a weekend	8

Table 2. Rule-based screening conditions and assigned risk contributions

The rule-based score is calculated by summing the risk contributions associated with all rules triggered by a transaction. Let $I_j(x)$ denote whether transaction x triggers rule j , where $I_j(x) = 1$ when the condition is satisfied and $I_j(x) = 0$ otherwise, and let w_j represent the corresponding rule weight. The unbounded rule score is therefore expressed as:

$$R^x(x) = \sum_{j=1}^7 w_j I_j(x)$$

Since multiple rules may be triggered simultaneously, the cumulative score can exceed the intended range of the rule-based component. PayGuard AI therefore caps the rule score at 80:

$$R(x) = \min(80, R^x(x))$$

where $R(x)$ represents the final rule-based risk score.

The assigned weights indicate the relative severity of control indicators within the prototype. Potential duplicate invoices and payments that exceed invoice amounts are assigned the highest individual contribution of 35 points, whereas contextual indicators, such as weekend payments, are assigned substantially lower contributions. These weights are based on prototype design decisions rather than empirically optimised estimates of fraud probability.

The rules differentiate transaction-level control indicators from contextual indicators. The payment-exceeds-invoice and invoice-exceeds-purchase-order rules use the nominal amount fields with thresholds of 1% and 5%, respectively; these screening conditions are separate from the currency-normalised exposure calculations in Section 3.7.1. Invoice numbers are normalised by converting them to uppercase and removing non-alphanumeric characters. Exact duplicates share the same supplier identifier, normalised invoice number and invoice amount rounded to two decimal places; the exact-match rule does not impose a date window. Near-duplicate comparisons are restricted to the same supplier and invoice dates no more than 60 days apart. A near match requires a SequenceMatcher similarity ratio of at least 0.90 but less than 1.00, and an absolute invoice-amount difference divided by the maximum of the two absolute invoice amounts and 1.0 of no more than 0.01. Both members of a matching pair receive duplicate-rule evidence. For exposure attribution, the exact-match

implementation selects records after the first member in batch order, while the near-match implementation selects the later member in invoice-date order. Recent bank-account changes and invoice/payment currency mismatches contribute separate control indicators. Rapid payment means a payment date zero or one day after the invoice date. Weekend payment receives the lowest contribution because weekend timing alone is contextual evidence rather than a planted anomaly. Multiple triggered conditions contribute to the combined rule score.

The rule-based score provides an interpretable measure of known control violations and suspicious payment characteristics. However, its effectiveness is limited by reliance on predefined conditions. To detect unusual transaction behaviour that may not be captured by explicit rules, PayGuard AI integrates Isolation Forest as a complementary unsupervised anomaly-detection mechanism.

3.5.2. Isolation Forest Anomaly Detection

The Isolation Forest model constitutes the second component of the hybrid detection mechanism. It is trained on the designated training partition and applied to the engineered numerical features. As an unsupervised algorithm, it does not utilise the `is_anomaly` labels during training; instead, it learns the underlying structure of the transaction data without access to the synthetic ground-truth labels identifying which observations were artificially introduced anomalies. The model subsequently assigns an anomaly score to each transaction. Its operational principle involves repeatedly generating random partitions by selecting a feature and a corresponding random split value, thereby separating observations and iterating this process. Typically, anomalous observations require fewer splits to be isolated and identified. Therefore,

$$\text{shorter isolation path} \Rightarrow \text{greater anomaly evidence}$$

Within the implemented framework, the Isolation Forest algorithm was configured with 200 estimators and a contamination parameter of 0.05. A fixed random state of 42 ensured reproducibility. The model was trained on eight engineered transaction features: paid-to-invoice ratio, invoice-to-purchase order ratio, days-to-pay, payment-term deviation, log-transformed payment amount, weekend-payment indicator, recent bank-change indicator, and currency-mismatch indicator. The eight-feature input matrix was transformed using a RobustScaler fitted exclusively to the training partition; the same fitted transformation was then applied to the held-out transactions.

Although contamination was set to 0.05, the framework uses percentile ranks of Isolation Forest's raw anomaly scores rather than its binary predictions. The contamination-dependent decision threshold therefore does not determine the hybrid alerts, which are governed by the rule contributions, hybrid weights and 45-point alert threshold.

After model training, the Isolation Forest algorithm generates a continuous anomaly score for each transaction. The output direction was adjusted so that higher values

indicate greater anomalousness. Since the raw Isolation Forest scores do not align with the 0–100 scale of the PayGuard AI risk framework, scores were converted to percentile ranks based on the training-score distribution. Each transaction thus receives an anomaly percentile reflecting its statistical deviation relative to the training data. This percentile-based transformation yields a bounded anomaly measure suitable for integration with the rule-based score in the hybrid risk assessment.

3.5.3. Hybrid Risk Score and Alert Generation

At this stage of the framework, two sources of detection evidence are utilised: the rule-based risk score and the Isolation Forest anomaly percentile, both described in Section 3.5.1 and Section 3.5.2 respectively. These metrics do not independently determine the final alert. The rule-based score provides explicit evidence of known payment-control violations, while the Isolation Forest percentile quantifies statistical deviation from the training data. These components are integrated to produce a unified hybrid risk score, enabling the system to incorporate both interpretable evidence from established fraud indicators and data-driven evidence of atypical behaviours that may elude predefined rules.

The hybrid risk score for transaction i is calculated as follows:

$$H_i = 0.72R_i + 0.28A_i$$

where H_i denotes the hybrid risk score, R_i represents the rule-based risk score, and A_i represents the Isolation Forest anomaly percentile. The rule-based component receives a higher weight because triggered business controls offer explicit and interpretable evidence of payment-risk conditions. In contrast, the Isolation Forest model provides complementary evidence of statistical anomalousness and is assigned a lower weight to prevent such anomalies from disproportionately influencing the operational risk assessment. The 0.72/0.28 weighting is a prototype design choice, not an empirically optimised parameter, and should be calibrated based on organisational risk tolerance, historical transaction data, and investigative capacity prior to operational deployment.

The hybrid score is subsequently converted into an operational alert using a predefined threshold of 45. A transaction is flagged when:

$$Alert_i = \begin{cases} 1, & H_i \geq 45 \\ 0, & H_i < 45 \end{cases}$$

where an alert value of 1 indicates that the transaction is prioritised for investigation. The 45-point threshold serves as a prototype operational choice rather than a universally applicable operational alert threshold and should be calibrated to align with an organisation's risk tolerance, anticipated investigation volume, and the relative costs of missed versus unnecessary alerts. Notably, the hybrid score is not interpreted as a calibrated probability of fraud. For instance, a hybrid score of 70 reflects a risk score of 70 on the framework's operational scale and does not correspond to a 70% probability that the transaction is fraudulent.

The Isolation Forest contribution is bounded above by 28 points because its percentile score cannot exceed 100 and its weight is 0.28. Consequently, a transaction with no triggered rules cannot reach the 45-point alert threshold. Even an unusually rapid payment triggering only the 18-point rapid-payment rule has a maximum hybrid score of $0.72 \times 18 + 0.28 \times 100 = 40.96$ and therefore cannot generate an alert. The implemented configuration uses statistical unusualness to modify ranking and alert decisions in the presence of sufficient rule-based evidence; it does not support alerts based solely on Isolation Forest evidence.

It is worth noting that PayGuard AI does not merely output a score; instead, its alert generation process retains information about the evidence contributing to transaction evaluation, rather than providing a binary classification. Investigators receive human-readable explanations that enable identification of explicit control conditions associated with the rule-based component, as well as the Isolation Forest percentile, which indicates the transaction's relative statistical unusualness. Consequently, the hybrid score and corresponding alert function as mechanisms for prioritising investigations, rather than serving as definitive fraud determinations. A flagged transaction signifies that the payment requires further review based on the combined assessment evidence, with the final determination made through human investigation.

3.6. Interpretable Statistical Modelling and Uncertainty Estimation

The hybrid detection mechanism identifies transactions requiring further investigation. Subsequently, a distinct statistical-analysis layer examines the relationships between transaction features and the planted anomaly outcome. This layer employs logistic regression to yield interpretable coefficients and odds-ratio estimates, as well as Bayesian logistic regression to quantify uncertainty in model parameters and predictive probabilities. The statistical modelling component complements the rule-based and Isolation Forest detection mechanisms, rather than serving as a substitute for primary transaction alert determination.

3.6.1. Logistic Regression Analysis

The logistic regression analysis employs the synthetic ground-truth `is_anomaly` label as the binary dependent variable, where $Y_i = 1$ denotes a planted anomaly and $Y_i = 0$ indicates a normal transaction. The independent variables consist of transaction characteristics derived through feature engineering. In contrast to the unsupervised Isolation Forest model, logistic regression leverages known synthetic labels to estimate the association between these features and the probability that a transaction is a planted anomaly. The primary objective is to provide interpretability and explanatory analysis, rather than to replace the existing hybrid detection mechanism.

For each transaction i , the logistic model represents the explanatory variables as a linear combination,

$$z_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$

where β_0 is the intercept, β_j represents the coefficient for feature j , and x_{ij} is the value of feature j for transaction i . The linear predictor is subsequently transformed into a probability using the logistic function,

$$P(Y_i = 1 | X_i) = \frac{1}{1 + e^{-z_i}}$$

Positive coefficients indicate that increases in the corresponding feature are associated with higher odds of planted anomaly status, while negative coefficients are associated with lower odds. Coefficients were estimated using L2-penalised logistic regression with $C = 1.0$, the lbfgs solver, and a maximum of 2,000 iterations.

Each logistic regression coefficient is also expressed as an odds ratio to facilitate interpretation of the estimated relationships,

$$OR_j = e^{\beta_j}$$

An odds ratio greater than 1 indicates that an increase in the corresponding predictor is associated with higher odds of planted anomaly status, whereas an odds ratio below 1 indicates lower odds. An odds ratio of 1 signifies no change in the odds associated with that predictor. Odds ratios thus offer a more interpretable representation of feature associations than raw log-odds coefficients. These relationships are interpreted as statistical associations and do not imply causality between transaction characteristics and planted anomaly status.

Because the continuous predictors were standardised using training-set statistics, their odds ratios correspond to a one-training-standard-deviation increase in the original engineered feature, holding the other model predictors fixed. Odds ratios for binary indicators compare an indicator value of 1 with 0.

Uncertainty in the logistic-regression coefficients was assessed using 1,000 nonparametric pairs-bootstrap resamples with random seed 7. Each resample contained 3,920 training transaction rows drawn with replacement, retaining each row's predictor values and anomaly label together. Continuous-feature scaling parameters estimated from the original training partition were held fixed across resamples. Each bootstrap model used L2-penalised logistic regression with $C = 1.0$, the lbfgs solver, and a maximum of 1,000 iterations. The 2.5th and 97.5th percentiles of the bootstrap coefficient distributions defined the 95% percentile intervals; exponentiating these endpoints produced the corresponding odds-ratio intervals. Resamples containing only one outcome class, fits raising a ValueError, fits producing a convergence warning, and fits returning non-finite coefficients were rejected and replaced, subject to a maximum of 3,000 attempts. The audited run accepted all 1,000 fits without rejecting any attempts. No explicit separation test was performed within individual resamples; L2 regularisation was retained to stabilise estimation in the presence of strongly separating predictors.

The logistic regression model is evaluated using both statistical and predictive measures. Coefficient estimates, odds ratios, and bootstrap uncertainty intervals are

examined to assess individual feature relationships. Overall model fit is further assessed using McFadden's pseudo- R^2 ,

$$R_{McFadden}^2 = 1 - \frac{\ln L_{model}}{\ln L_{null}}$$

where L_{model} is the Bernoulli likelihood evaluated at the fitted penalised-model coefficients and L_{null} is the likelihood of the intercept-only model evaluated on the same training observations. The reported value is a descriptive likelihood-based fit measure for the penalised model and should not be interpreted as the proportion of variance explained. Predictive discrimination is separately evaluated on held-out data using ROC-AUC and average precision.

The hybrid score was compared with its constituent rule-based and Isolation Forest scores, as well as logistic-regression predicted probabilities, using the same held-out transactions. ROC-AUC and average precision were calculated from un-thresholded numerical scores rather than binary alert decisions. The rule-only comparator used the existing rule weights without additional fitting or tuning.

3.6.2. Bayesian Logistic Regression and Uncertainty Estimation

The L2-penalised logistic-regression analysis provides coefficient point estimates with uncertainty summarised by the bootstrap intervals described in Section 3.6.1. Bayesian logistic regression provides a different representation of parameter uncertainty through a posterior distribution conditional on the specified likelihood and priors. In this study, a variational approximation to that posterior was used to summarise uncertainty in feature associations and propagate parameter uncertainty into transaction-level predicted anomaly probabilities.

Bayesian estimation integrates prior assumptions regarding model parameters with evidence derived from observed training data. This relationship is expressed as

$$P(\beta \mid D) \propto P(D \mid \beta)P(\beta)$$

where $P(\beta \mid D)$ denotes the posterior distribution of the model parameters after observing data D , $P(D \mid \beta)$ denotes the likelihood of the observed data given the parameters, and $P(\beta)$ denotes the prior distribution. For transaction i , the binary planted-anomaly outcome is modeled as

$$Y_i \sim \text{Bernoulli}(p_i)$$

with the probability of planted anomaly status related to the transaction characteristics through

$$\text{logit}(p_i) = \alpha + x_i^\top \beta$$

where α denotes the intercept, $x_i^\top$ represents the transaction features, and β represents their corresponding regression coefficients. Prior distributions are specified for these

parameters, and the likelihood of the observed transaction data is then incorporated to obtain their posterior distributions. As a result, the analysis yields a distribution of plausible parameter values instead of a single coefficient estimate.

The Bayesian logistic regression model was fitted using six transaction features: log-transformed payment amount, paid-to-invoice ratio, invoice-to-PO ratio, days-to-pay, payment-term deviation, and the weekend-payment indicator. The recent bank-change and currency-mismatch indicators were excluded from this Bayesian specification because every training transaction exhibiting either indicator was a planted anomaly, creating a deterministic relationship between the indicator and the outcome. Their exclusion was a modelling choice intended to examine the remaining six features, rather than evidence that Bayesian estimation requires their removal. Their exclusion is limited to the Bayesian inferential model; both indicators remain incorporated in the rule-based, Isolation Forest, hybrid detection, and conventional logistic-regression analyses. Continuous predictors were standardised using the mean and standard deviation calculated exclusively from the training data, while the weekend indicator retained its 0/1 representation. Weakly informative Normal prior distributions were assigned to the model parameters, with the intercept specified as

$$\alpha \sim N(0, 2.5^2)$$

and each feature coefficient specified as

$$\beta_j \sim N(0, 1.5^2)$$

Posterior inference was approximated using Automatic Differentiation Variational Inference (ADVI) implemented in PyMC. Rather than obtaining posterior samples through Markov chain Monte Carlo, ADVI optimises a mean-field variational approximation to the posterior distribution. Optimisation was performed for a maximum of 30,000 iterations using a fixed random seed of 42, with convergence monitored through absolute changes in the variational parameters. Following optimisation, 4,000 draws were generated from the fitted approximate posterior and used to summarise coefficient distributions and transaction-level predictive uncertainty. The optimisation loss was retained to assess the progression of the variational fitting procedure.

The resulting approximate posterior distributions allow direct examination of uncertainty surrounding individual model parameters. Posterior summaries include central coefficient estimates and 95% credible intervals. The reported equal-tailed 95% intervals contain the central 95% probability mass of the fitted variational posterior approximation, conditional on the specified model, priors and training data; their accuracy also depends on the adequacy of that approximation. The position and width of these intervals provide information about both the estimated association and its uncertainty. For instance, a narrow posterior interval suggests greater precision in the parameter estimate, while a wider interval indicates greater uncertainty.

Parameter uncertainty is also propagated into transaction-level predictions. For each held-out transaction, posterior coefficient samples generate a distribution of predicted anomaly probabilities rather than a single deterministic estimate. These posterior predictive probabilities are summarised by their central estimate and corresponding credible bounds, enabling the framework to quantify uncertainty in individual transaction assessments. Narrower credible intervals for predicted probabilities indicate greater certainty under the fitted model, while wider intervals suggest that the model and data provide less precise evidence regarding the transaction's anomaly status. This information can assist human reviewers by distinguishing confident model assessments from those requiring greater caution.

The Bayesian probabilities and credible intervals do not replace the primary hybrid alert mechanism described in Section 3.5. Instead, they offer an additional uncertainty-aware analytical layer that complements the interpretable associations provided by conventional logistic regression. Because the Bayesian model is trained on synthetic ground-truth labels, its predictive probabilities reflect estimated probabilities of planted anomaly status under the specified model and should not be interpreted as calibrated probabilities of real-world fraud. Thus, Bayesian logistic regression enhances PayGuard AI by explicitly representing uncertainty in both model relationships and transaction-level predictions, supporting more transparent interpretation of analytical results.

The original Bayesian fit used PyMC 5.26.1, with ArviZ 0.23.4 recorded in the saved posterior metadata. Inference used mean-field Automatic Differentiation Variational Inference (ADVI), with a maximum of 30,000 optimisation iterations and random seed 42. A parameter-convergence callback checked absolute changes every 100 iterations using a tolerance of 10^{-4} ; the recorded run completed all 30,000 iterations. Following optimisation, 4,000 draws were generated from the variational posterior using random seed 42. Approximate 95% credible intervals were calculated using the 2.5th and 97.5th percentiles of these draws. The posterior draws and optimisation-loss history were saved for subsequent analysis and figure generation.

3.7. Financial Risk Analysis and Decision Support

Earlier stages of the framework utilise a hybrid detection mechanism to identify transactions requiring further investigation and statistical modelling, offering interpretations and explanations of uncertainty. However, the resulting hybrid risk score does not quantify the monetary significance of these findings or address the financial implications of alternative payment strategies. Therefore, the final layer of PayGuard AI integrates two complementary financial analyses. The first estimates review exposure for flagged transactions and applies bootstrap resampling to assess aggregate exposure uncertainty. The second evaluates working-capital outcomes under various Days Payable Outstanding (DPO) extensions, employing Monte Carlo simulation to account for uncertain financial assumptions. These analyses serve as proof-of-concept decision support tools and should not be regarded as calibrated forecasts of organisational fraud losses or treasury outcomes.

3.7.1. Review Exposure Estimation

For each alerted transaction, a review-exposure measure is calculated to provide a monetary context alongside the hybrid risk score. The hybrid risk score quantifies the combined evidence from rule-based and anomaly-detection methods, whereas the review-exposure measure indicates the monetary value associated with the transaction that may require further investigation. This measure is not interpreted as a confirmed or expected loss, as flagged transactions may ultimately be legitimate. Rather, it serves as an operational prioritisation tool, enabling investigators to assess both the analytical strength and the potential financial impact of an alert.

When a monetary discrepancy is directly observable, the exposure measure is derived from transaction values. In the case of an overpayment, the measurable exposure is calculated as follows:

$$E_i^{\text{over}} = \max(0, P_i - I_i)$$

where P_i denotes the paid amount converted to SMU using the payment-currency factor, and I_i denotes the corresponding invoice amount converted to SMU using the invoice-currency factor. Each converted amount is rounded to two decimal places before the exposure calculation. For currency-mismatch cases, the resulting difference is conditional on the fixed illustrative conversion assumptions.

Consequently, only the amount paid above the invoiced obligation is treated as overpayment exposure rather than the full payment value. For potential duplicate transactions, exposure is attributed to the later transaction identified as the duplicate candidate rather than to both members of the matching pair. The candidate's full paid amount is treated as duplicate-candidate exposure. This prevents the same potential duplicate event from being counted twice while retaining the payment value that may represent a repeated financial obligation.

A transaction may simultaneously exhibit more than one directly measurable exposure condition. To reduce the possibility of double-counting overlapping monetary exposure, PayGuard AI does not sum the overpayment and duplicate-candidate amounts. Instead, direct exposure is defined as the larger of the two values:

$$E_i^{\text{direct}} = \max(E_i^{\text{over}}, E_i^{\text{dup}})$$

where E_i^{dup} denotes the duplicate-candidate exposure. This method ensures that the directly measurable review exposure reflects the largest observable monetary amount associated with the alert, without presuming that overlapping indicators represent distinct financial losses. However, some alerts do not involve a directly observable monetary discrepancy. For instance, a transaction may be flagged due to unusual payment timing, contextual control indicators, or statistical deviations identified by Isolation Forest, without an identifiable duplicate or overpayment amount. For such

transactions lacking measurable direct exposure, PayGuard AI calculates an illustrative scenario exposure as follows:

$$E_i^{\text{scenario}} = P_i \left(\frac{H_i}{100} \right) (0.25)$$

where P_i denotes the paid amount converted to the same SMU base and H_i denotes the hybrid risk score. The factor of 0.25 is a prototype scenario assumption, not an empirically derived loss-given-fraud parameter. Thus, the scenario calculation offers a monetary prioritisation value for alerts without directly measurable exposure and should not be interpreted as an estimate of expected fraud loss.

The final review exposure assigned to an alerted transaction is therefore defined as follows:

$$E_i^{\text{review}} = \begin{cases} E_i^{\text{direct}}, & E_i^{\text{direct}} > 0 \\ E_i^{\text{scenario}}, & E_i^{\text{direct}} = 0 \end{cases}$$

Directly observable exposure is prioritised whenever available, while scenario exposure offers an illustrative monetary measure when no such discrepancy can be calculated. This distinction maintains the separation between measurable transaction discrepancies and assumption-based financial estimates. The resulting review-exposure values are then used in bootstrap analysis to assess uncertainty surrounding aggregate exposure across held-out alerted transactions.

3.7.2. Bootstrap Analysis of Aggregate Review Exposure

Bootstrap resampling is applied to the held-out alerted transactions to evaluate sampling variability in aggregate review exposure. When the set of held-out alerted transactions contains n review-exposure observations, each bootstrap iteration samples n observations with replacement from this empirical distribution. The aggregate review exposure for bootstrap iteration b is then computed as

$$T^{*(b)} = \sum_{i=1}^n E_i^{*(b)}$$

This procedure permits transactions to appear multiple times or not at all within a given resample, thereby generating alternative samples that remain consistent with the observed held-out alert population.

The bootstrap procedure comprised 20,000 resamples with random seed 42. The sampling unit was an individual flagged held-out transaction. Each resample contained 66 transaction-level review-exposure values drawn with replacement from the 66 flagged transactions, and these values were summed. The resulting distribution was summarised by its mean, median, 5th, 95th, and 99th percentiles, together with the conditional mean at or above the 95th percentile. The 5th and 95th percentiles delimit a central 90% bootstrap resampling range, not a 95% confidence interval. These

summaries are conditional on the observed flagged set and specified exposure assumptions and do not represent calibrated fraud-loss VaR or CVaR estimates. No regression models were fitted within this bootstrap, so failed-fit and separation handling were not applicable.

3.7.3. Monte Carlo Working-Capital Simulation

A separate Monte Carlo analysis assesses the potential working-capital implications of alternative supplier-payment timing strategies. Days Payable Outstanding (DPO) is defined as the average period during which an organisation retains cash before settling supplier obligations. Extending DPO may provide a working-capital benefit by allowing cash to remain available to the organisation for a longer duration. However, delaying supplier payments can also introduce financial costs, such as the loss of early-payment discounts and increased supplier-friction costs. Due to the uncertainty in the magnitude of these benefits and costs, PayGuard AI applies Monte Carlo simulation to generate distributions of possible financial outcomes instead of relying on a single deterministic estimate.

For this proof-of-concept analysis, annual supplier spend is estimated using transactions identified as normal by the synthetic ground-truth labels. This constitutes an oracle assumption that is possible because the experiment uses controlled synthetic data in which the true planted-anomaly status of each transaction is known. In an operational environment, the true status of suspicious transactions would not be known with certainty before investigation. The total value of normal transactions is therefore annualised according to the period represented by the synthetic dataset. The number of years represented is calculated as

$$Y = \frac{\max\{(D_{\max} - D_{\min})_{\text{days}} + 1, 1\}}{365.25}$$

where $D_{\max}$ and $D_{\min}$ are the latest and earliest payment dates among transactions labelled normal. The additional day makes the observation span inclusive of both endpoints, and the maximum ensures a minimum span of one day. Annualised supplier spend is subsequently calculated as

$$S = \frac{\sum_{i=1}^{n_{\text{normal}}} P_i}{Y}$$

where n_{normal} is the number of normal-labelled transactions and P_i is the paid amount of normal transaction i converted to the common SMU base using its payment-currency factor.

Three alternative payment-timing scenarios are evaluated, corresponding to DPO extensions of 10, 20, and 30 days:

$$\Delta\text{DPO} \in \{10, 20, 30\}.$$

For each scenario, the additional working capital retained through delayed payment is approximated as

$$WC_{\Delta} = S \frac{\Delta DPO}{365},$$

where S represents annualised supplier spend. The financial benefit associated with retaining this working capital is then calculated as

$$B_{\Delta} = WC_{\Delta} r_c,$$

where r_c represents the simulated annual cost of capital. A higher cost of capital therefore increases the estimated financial benefit of retaining cash for a longer period.

The financial benefit of extending DPO is evaluated alongside two potential costs: foregone early-payment discounts and supplier-friction costs. Because the precise values of the relevant financial parameters are uncertain, probability distributions are specified for each parameter and sampled during the Monte Carlo simulation. Table 3 presents the distributions used in the implemented prototype. These distributions represent illustrative modelling assumptions and were not calibrated using treasury data from a specific organisation.

Input parameter	Distribution	Representation in the model
Cost of capital (r_c)	Normal (mean = 9%, SD = 1.5%), clipped to 3%–18%	Annual financing cost used to value retained working capital
Discount-capture share (q_d)	Beta(6, 14)	Proportion of supplier spend currently capturing early-payment discounts
Early-payment discount rate (r_d)	Normal (mean = 2%, SD = 0.4%), clipped to 0.5%–5%	Financial discount currently obtained on eligible supplier spend
Discount erosion (e_d)	Beta(4, 6)	Proportion of existing discount benefits assumed to be lost following the DPO extension
Supplier-friction exposure (q_f)	Beta(2, 40)	Proportion of supplier spend exposed to potential friction costs
Supplier-friction cost rate (r_f)	Normal (mean = 1.5%, SD = 0.5%), clipped to 0%–5%	Financial cost associated with supplier friction

Table 3. Probability distributions used in the Monte Carlo working-capital simulation

The six inputs were sampled independently. Their distributions were held constant across the 10-, 20- and 30-day scenarios; only the financing-benefit term depended on the length of the DPO extension.

In each simulation, the financial cost resulting from lost early-payment discounts is calculated as follows:

$$L_{\text{discount}} = Sq_d r_d e_d,$$

where q_d denotes the simulated proportion of spend currently capturing discounts, r_d denotes the simulated discount rate, and e_d denotes the simulated proportion of those benefits lost after the DPO extension. The supplier-friction cost is calculated in a similar manner as

$$C_{\text{friction}} = Sq_f r_f,$$

where q_f denotes the simulated proportion of supplier spend exposed to friction and r_f denotes the corresponding simulated financial cost rate. The net financial benefit for each DPO scenario is then calculated as

$$NB_{\Delta} = B_{\Delta} - L_{\text{discount}} - C_{\text{friction}}.$$

This calculation allows each simulated outcome to incorporate both the financing benefit of retaining working capital and the potential financial costs associated with changes in supplier-payment timing.

Each of the 10-, 20-, and 30-day DPO scenarios was evaluated using 20,000 Monte Carlo iterations. Simulation continued the random-number generator initialised with seed 42 for the preceding review-exposure bootstrap; the seed was not reset before the DPO scenarios. The resulting net-benefit distributions are summarised using the mean, 10th percentile, median, and 90th percentile. Additionally, the proportion of simulations in which $NB_{\Delta} > 0$ is calculated to indicate the frequency with which each DPO scenario yields a positive net financial outcome under the specified model assumptions. This measure reflects the proportion of positive outcomes within the simulated scenario and should not be interpreted as an empirically calibrated probability that a DPO strategy would be profitable in an actual organisation.

The Monte Carlo analysis extends PayGuard AI beyond transaction-level fraud screening by demonstrating how uncertainty can be incorporated into treasury and working-capital decision support. Instead of recommending a specific DPO extension based on a single estimated outcome, the simulation provides a distribution of potential financial consequences for each payment-timing scenario. However, because the analysis relies on synthetic transaction data, ground-truth normal transaction labels, and illustrative financial assumptions, the results should be interpreted as a proof-of-concept demonstration of the decision-support methodology rather than as organisation-specific financial forecasts.

4. Results

4.1. Experimental Setup and Dataset Composition

The final synthetic dataset consisted of 5,600 supplier-payment transactions. Of these, 5,320 represented standard payment behaviour, while the remaining 280 were intentionally planted anomalies, resulting in an anomaly prevalence of 5%. The anomalies were categorized into seven types: exact duplicate, near duplicate, overpayment, invoice-above-PO, recent bank-account change, currency mismatch, and rapid payment, with 40 observations in each category. The `is_anomaly` label denoted synthetic anomaly scenarios and did not indicate confirmed fraud cases. Fig. 2 presents the composition of the dataset.

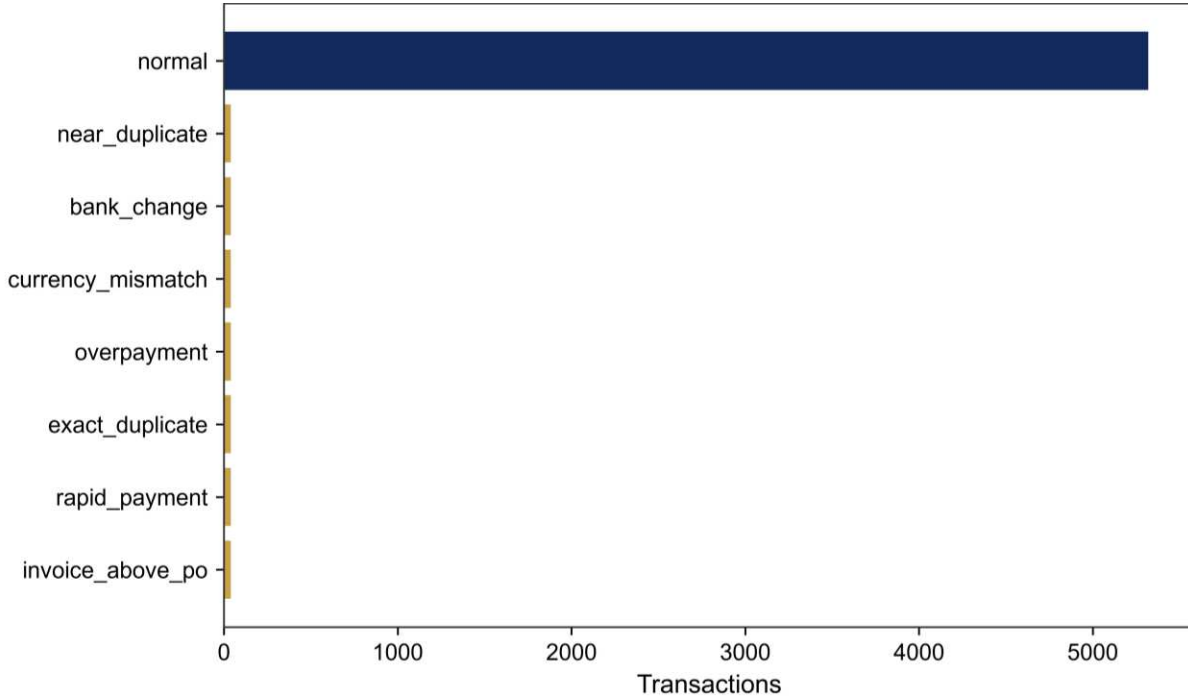


Fig. 2 Composition of the synthetic supplier-payment dataset by transaction class and planted anomaly category

For experimental evaluation, the dataset was partitioned into 70% training observations (3,920 transactions) and 30% held-out test observations (1,680 transactions). The split was stratified by planted anomaly status and used random seed 42. The training transactions were used to derive preprocessing parameters and fit the model. Subsequently, the fitted model was applied to the held-out test transactions to assess performance. This approach ensured that the observations used for final evaluation did not influence model fitting.

4.2. Hybrid Fraud-Detection Performance

Evaluation on the held-out test set demonstrated that the hybrid PayGuard detector achieved a precision of 0.848, recall of 0.667, and F1-score of 0.747. The specificity was 0.994, corresponding to a false-positive rate of 0.006. The continuous hybrid risk score yielded an ROC-AUC of 0.982 and an average precision of 0.770. At the selected alert threshold, 3.93% of held-out transactions were flagged for investigation, which is

equivalent to approximately 39.29 alerts per 1,000 transactions. Table 4 presents the principal performance measures for the held-out set.

Metric	Final value
Precision	0.848
Recall	0.667
F1-score	0.747
Specificity	0.994
False-positive rate	0.006
ROC-AUC	0.982
Average precision	0.770
Alert rate	3.93%
Alerts per 1,000	39.29

Table 4. Held-out hybrid detection performance

Table 5 summarises the comparative performance. The rule score alone achieved the highest ROC-AUC on the held-out synthetic dataset (0.989), followed by the hybrid score (0.982), logistic regression (0.889), and Isolation Forest alone (0.826). However, the hybrid score achieved higher average precision than the rule score alone (0.770 versus 0.680). These findings indicate a metric-dependent trade-off: the hybrid improved average precision but did not improve ROC-AUC relative to the rule-based component. The comparisons are descriptive; statistical significance of the differences was not assessed.

Detection score/model	ROC-AUC	Average precision
Rule score alone	0.989	0.680
Hybrid risk score	0.982	0.770
Logistic regression	0.889	0.766
Isolation Forest alone	0.826	0.629

Table 5. Comparison of un-thresholded detection scores on the held-out test set

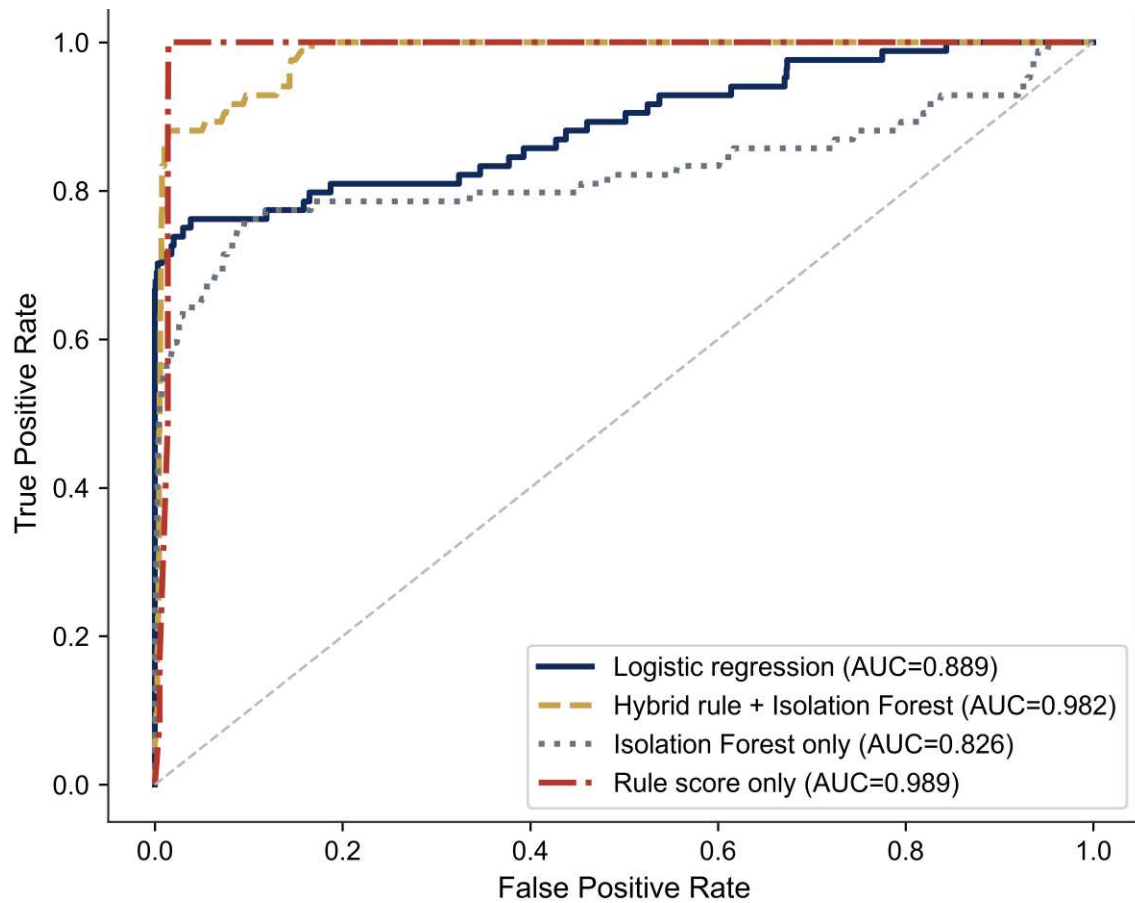


Fig. 3 ROC curves for the hybrid risk score, rule score alone, logistic regression, and Isolation Forest on the held-out test set

The confusion matrix presented in Fig. 4 offers additional insight into model performance at the selected alert threshold. Of the 84 planted anomalies in the held-out partition, 56 were correctly flagged, while 28 remained below the alert threshold. Among the 1,596 normal transactions, 1,586 were correctly left unflagged, and 10 resulted in false-positive alerts. The resulting precision of 84.8% demonstrates that most generated alerts corresponded to planted anomaly cases. In contrast, the recall of 66.7% reveals that approximately one-third of the planted anomalies were not identified at the selected operating threshold.

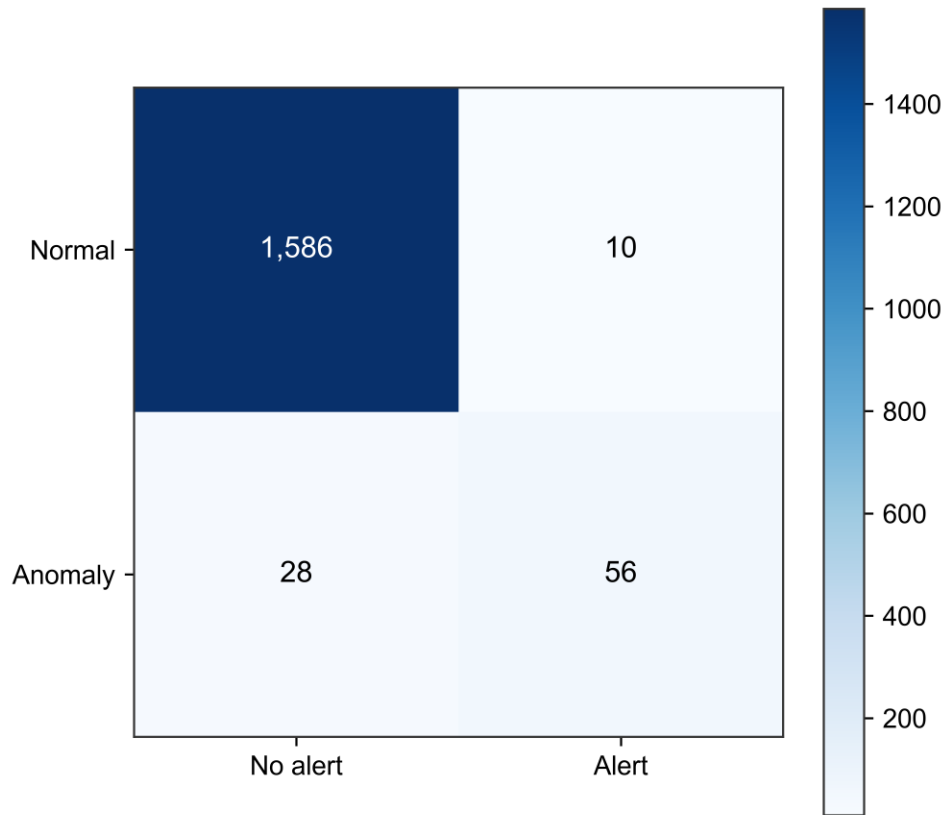


Fig. 4 Confusion matrix of the hybrid detector on the held-out test set

Detection performance differed markedly across the planted anomaly categories. As illustrated in Fig. 5, all held-out bank-change, currency-mismatch, and overpayment anomalies were detected, resulting in detection rates of 100% for each category. Invoice-above-PO anomalies were also identified with high consistency, with 15 of 17 cases detected (88.2%). In contrast, performance was lower for rapid-payment anomalies, with 4 of 13 detected (30.8%), and for near duplicates, with 2 of 13 detected (15.4%). None of the six held-out exact-duplicate cases were flagged at the selected hybrid threshold. These findings indicate that aggregate performance metrics obscure significant variation in the framework's sensitivity to different types of anomalous payment behaviour.

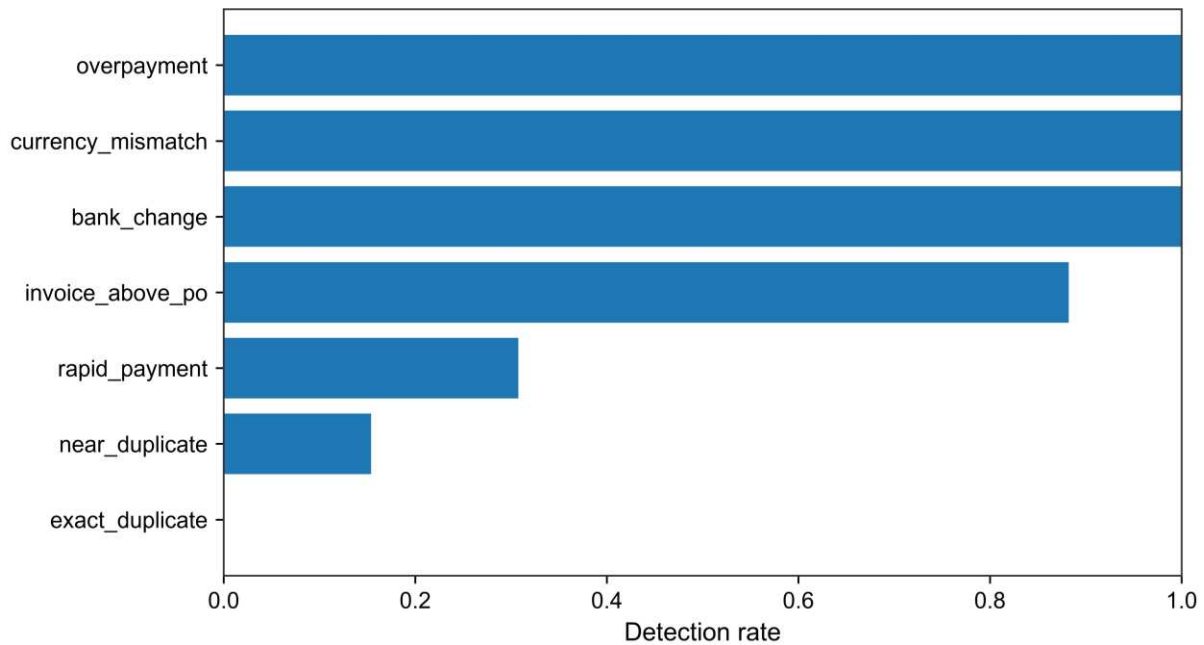


Fig. 5 Detection rates of the hybrid framework across planted anomaly categories in the held-out test set

4.3. Logistic Regression Analysis

The L2-penalised logistic regression model achieved a receiver operating characteristic area under the curve (ROC-AUC) of 0.889 and an average precision of 0.766 on the held-out test set. The model also produced a McFadden pseudo- R^2 of 0.469 during the training analysis. These findings suggest that the engineered payment features contained substantial information relevant to planted anomaly status. While logistic regression demonstrated lower held-out discrimination compared to the hybrid risk score, its primary function within PayGuard AI was to provide an interpretable statistical representation of feature associations, rather than to serve as the main detection mechanism.

Fig. 6 displays the estimated feature effects as odds ratios. Currency mismatch and recent bank account change exhibited the largest estimated associations with planted anomaly status, with odds ratios of approximately 340.22 and 320.83, respectively. These exceptionally large values result from the controlled synthetic construction of the corresponding anomaly scenarios and should not be interpreted as estimates of real-world fraud risk. Among the continuous engineered features, a one-training-standard-deviation increase in the paid-to-invoice ratio was associated with an odds ratio of 5.58, while the corresponding odds ratio for the invoice-to-purchase-order ratio was 1.97. A one-training-standard-deviation increase in payment-term deviation was associated with an odds ratio of 0.47, holding the other predictors fixed.

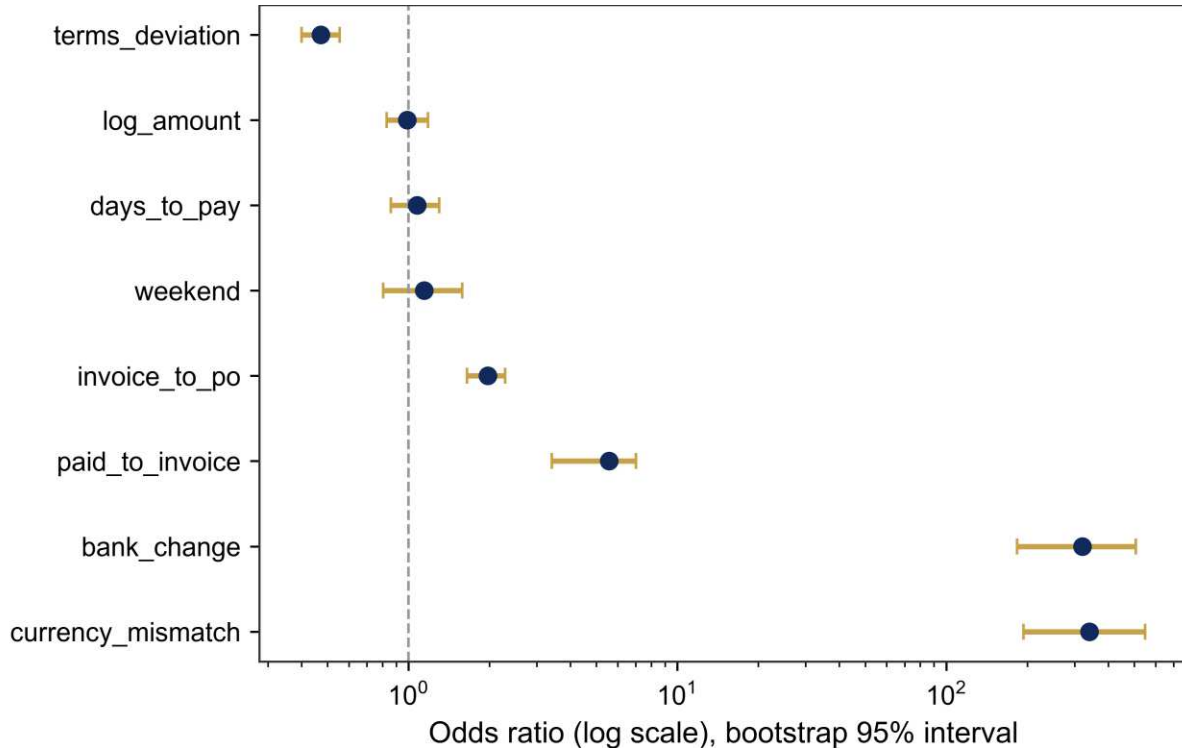


Fig. 6 Estimated odds ratios and 95% bootstrap intervals from the L2-penalised logistic-regression analysis. Continuous-feature effects are expressed per training-set standard deviation; binary-feature effects compare 1 with 0

Several features demonstrated limited evidence of a distinct association within the synthetic dataset. The bootstrap interval for the weekend-payment indicator ranged from 0.81 to 1.58. The intervals for days-to-pay and log payment amount ranged from 0.86 to 1.30 and 0.83 to 1.18, respectively. Each of these intervals included an odds ratio of one. In contrast, the intervals for the paid-to-invoice ratio, invoice-to-PO ratio, and payment-term deviation were entirely on one side of one, indicating more clearly differentiated associations with planted anomaly status under the experimental conditions.

4.4. Bayesian Uncertainty Analysis

The Bayesian logistic-regression model was estimated using mean-field automatic differentiation variational inference (ADVI), which completed the maximum of 30,000 optimization iterations. Subsequently, 4,000 draws were generated from the fitted approximate posterior distribution. On the held-out test set, the model achieved an ROC-AUC of 0.748, an average precision of 0.543, and a Brier score of 0.033. These metrics describe the reduced six-feature Bayesian specification. They do not isolate the effect of Bayesian rather than conventional estimation because the models use different predictor sets: the Bayesian specification excludes the bank-change and currency-mismatch indicators that are deterministically linked to planted anomaly status in the training data. Because mean-field ADVI provides a variational approximation rather than

exact posterior sampling, the resulting posterior distributions and credible intervals should be interpreted as approximate uncertainty estimates. Mean-field variational inference may underrepresent posterior dependence and uncertainty when relationships among model parameters are not fully captured.

Fig. 7 presents the approximate posterior coefficient distributions. The paid-to-invoice ratio demonstrated the strongest positive posterior association with planted anomaly status, with a posterior mean coefficient of 2.120 and an approximate 95% credible interval ranging from 0.876 to 3.368. The invoice-to-purchase order (PO) ratio also showed a consistently positive association, with a posterior mean of 0.535, an approximate 95% credible interval from 0.405 to 0.659, and an effectively 100% posterior probability of a positive effect. In contrast, payment-term deviation yielded a posterior mean of -0.557 , with its approximate 95% credible interval entirely below zero (-0.691 to -0.426). The intervals for log payment amount, days-to-pay, and the weekend indicator included zero, indicating greater uncertainty regarding their directional associations within the reduced Bayesian specification.

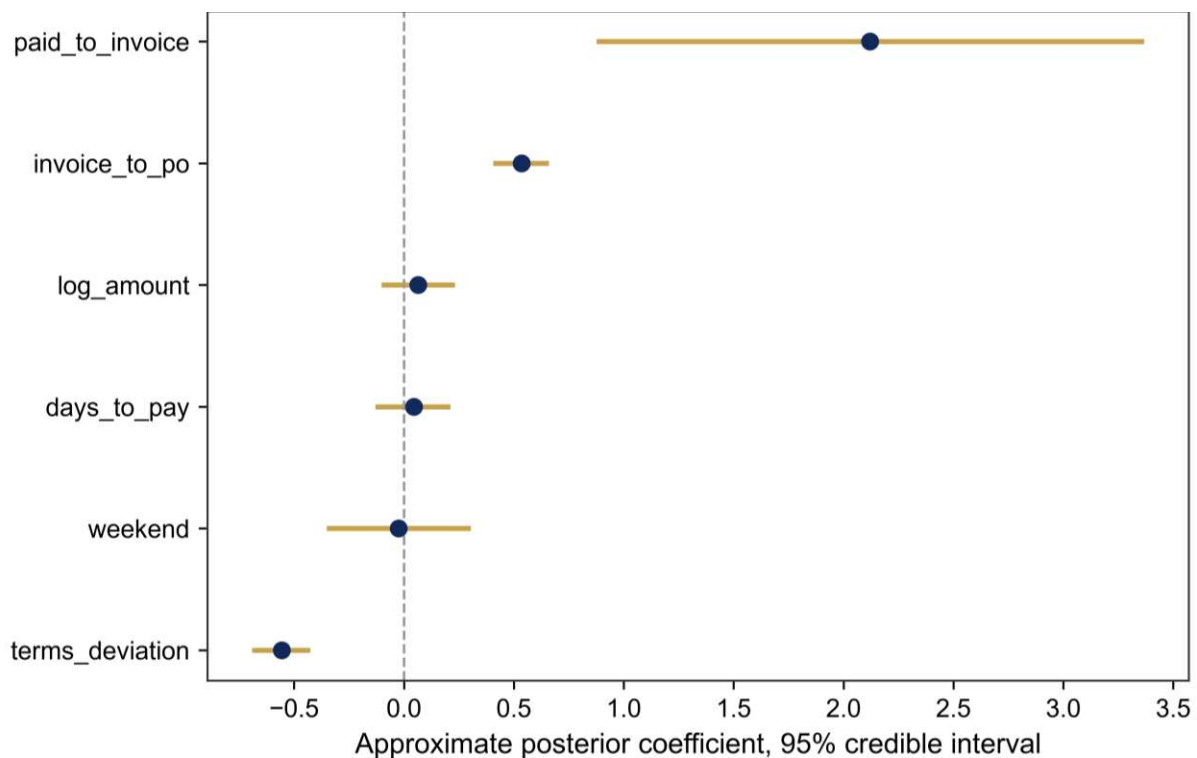


Fig. 7 Approximate posterior coefficient means and 95% credible intervals for the Bayesian logistic-regression model. Continuous-feature coefficients are expressed per training-set standard deviation; the weekend coefficient compares 1 with 0

4.5. Review-Exposure Analysis

A bootstrap analysis was conducted on the review-exposure values for the 66 transactions flagged in the held-out test set, utilising 20,000 resamples. The aggregate

exposure distribution produced a mean of approximately 302,564 SMU and a median of 286,918 SMU. The 5th and 95th percentiles were approximately 147,400 SMU and 514,441 SMU, respectively, and the 99th percentile reached 634,443 SMU. For simulations at or above the 95th percentile, the conditional mean aggregate exposure was approximately 587,257 SMU. These results reflect review exposure under repeated resampling of the flagged transaction set, rather than confirmed or predicted fraud losses.

Fig. 8 presents the distribution of aggregate review exposure. The variability observed in the bootstrap outcomes indicates that the financial magnitude associated with the flagged transaction set is influenced by sampling variability and cannot be fully captured by a single point estimate. Consequently, upper-percentile measures complement the mean exposure by highlighting the magnitude of higher-exposure resampling outcomes.

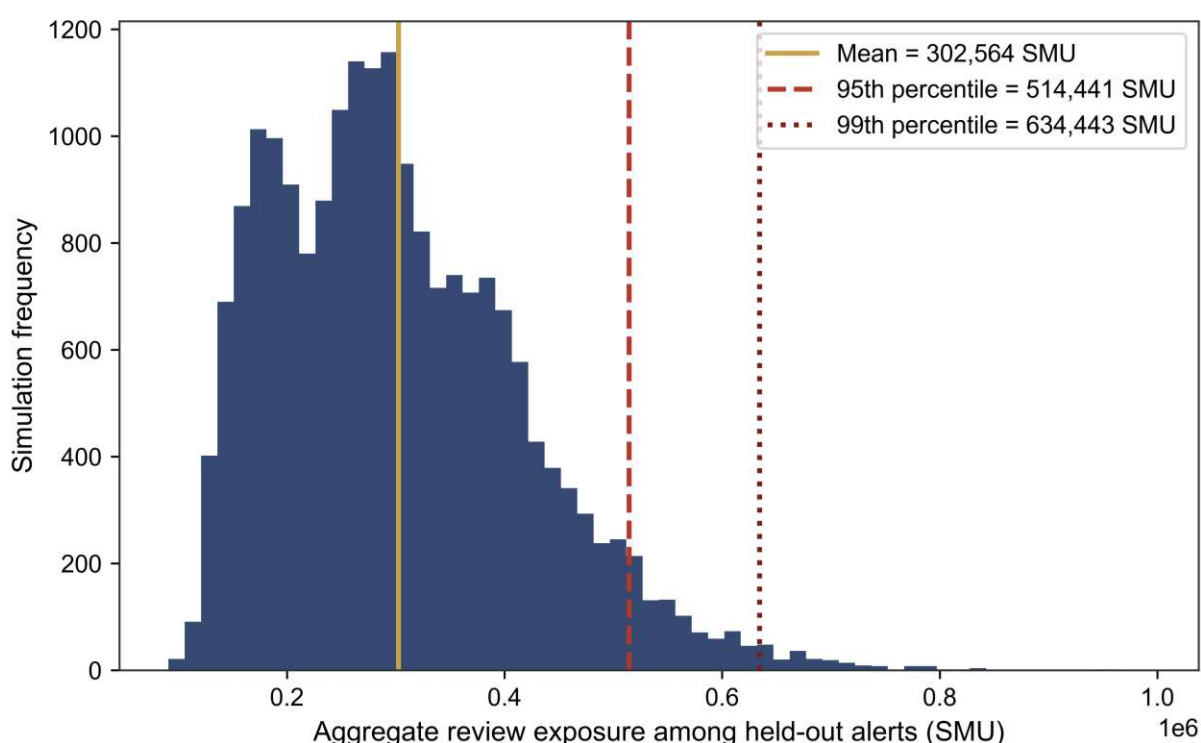


Fig. 8 Bootstrap distribution of aggregate review exposure across 20,000 resamples of flagged held-out transactions (SMU)

4.6. Monte Carlo Working-Capital Analysis

The working-capital analysis employed 20,000 Monte Carlo simulations to assess 10-, 20-, and 30-day extensions to Days Payable Outstanding, accounting for uncertainty in financing benefits, discount erosion, and supplier-friction costs. The synthetic normal-payment data reflected an annualised spend of approximately 38.78 million SMU. As detailed in Section 3.7, these values constitute an illustrative synthetic scenario and are not tailored to the treasury position or supplier relationships of any specific organisation.

The simulated financial outcomes differed across the three Days Payable Outstanding extension scenarios. A 10-day extension resulted in a mean net benefit of –25,171 SMU and a median of –16,382 SMU, with only a 37.2% probability of achieving a positive net outcome. In contrast, a 20-day extension yielded a mean net benefit of 71,310 SMU and a median of 78,449 SMU, with the probability of a positive outcome rising to 86.7%. The 30-day extension generated the highest simulated mean net benefit at 166,905 SMU and a median of 172,575 SMU, with a 97.8% probability of a positive outcome.

The distributional results presented in Fig. 9 further highlight the distinctions among the scenarios. For the 10-day extension, the 10th and 90th percentiles ranged from approximately –102,621 SMU to 41,928 SMU. The 20-day scenario exhibited a range from –14,376 SMU to 147,256 SMU, while the 30-day scenario ranged from 71,272 SMU to 256,042 SMU. Under the assumptions of the synthetic simulation, the 30-day scenario remained positive even at the 10th percentile, whereas the lower tails of the 10- and 20-day scenarios included negative outcomes. These results illustrate that the decision-support layer can capture both expected financial outcomes and the uncertainty associated with alternative working-capital scenarios, rather than relying exclusively on deterministic point estimates.

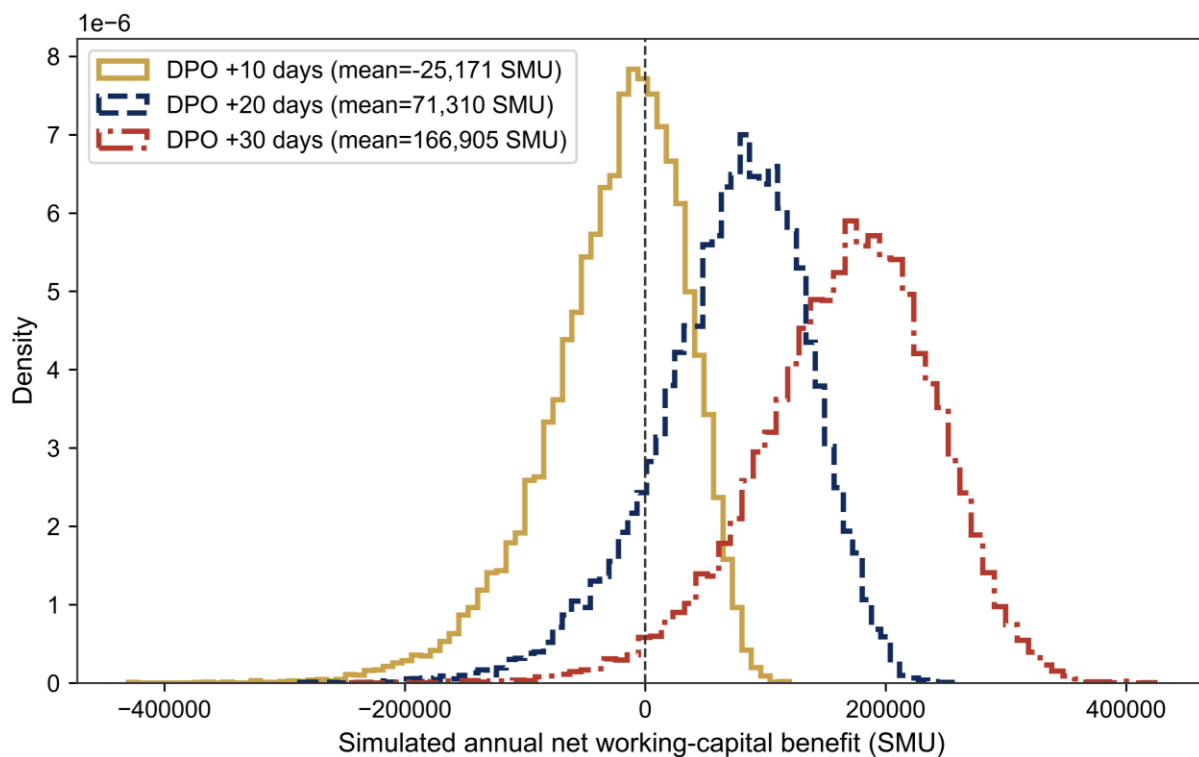


Fig. 9 Monte Carlo distributions of net financial benefit, expressed in SMU, for 10-, 20-, and 30-day DPO-extension scenarios

5. Discussion

This section analyses the previously reported experimental findings in relation to the three research questions that informed the development of the PayGuard AI framework. The results indicate that integrating hybrid detection, interpretable statistical modelling, Bayesian uncertainty estimation, and financial decision-making support yields a unified Accounts Payable decision-support framework. Additionally, analysis of the data reveals substantial variation in detection performance across different anomaly categories and underscores the limitations inherent in the controlled synthetic experimental design. Therefore, these findings are best interpreted as evidence regarding the behaviour and feasibility of the proposed proof-of-concept framework, rather than as predictors of expected performance in an actual corporate Accounts Payable setting.

5.1. Hybrid Detection and RQ1

The findings address RQ1 by demonstrating strong overall discrimination under controlled synthetic conditions, while showing that the hybrid score does not uniformly outperform its constituent components. The hybrid achieved a ROC-AUC of 0.982, exceeding those of logistic regression (0.889) and Isolation Forest alone (0.826), but falling below the rule-only score (0.989). Conversely, its average precision exceeded the rule-only baseline (0.770 versus 0.680). The contribution of the combined score therefore depends on the evaluation criterion. At the selected operating threshold, the hybrid achieved a precision of 84.8% and specificity of 99.4%, with a false-positive rate of approximately 0.6%. Further evaluation at comparable review capacities is needed to establish whether the hybrid offers practical advantages over rule-based screening.

Performance differed across anomaly categories. At the selected hybrid threshold, 56 of 84 planted anomalies were flagged, giving a recall of 66.7%. All bank-change, currency-mismatch and overpayment cases were flagged, while invoice-above-PO and rapid-payment sensitivity were 88.2% and 30.8%, respectively. The duplicate rule produced evidence for all six held-out exact duplicates and all thirteen held-out near duplicates, but the final hybrid alert flagged none of the exact duplicates and only two near duplicates (15.4%). The low duplicate alert sensitivity therefore reflects the combined scoring and threshold configuration on these cases, rather than failure of the matching rule to recognise the pairs. High overall discrimination consequently does not imply uniform alert sensitivity across anomaly categories.

These findings demonstrate the feasibility of the hybrid design but do not establish uniform superiority over rule-based screening or comprehensive anomaly coverage. The individual rule weights, the 0.72/0.28 weighting between the rule and Isolation Forest components, and the 45-point alert threshold were prototype design choices rather than empirically optimised settings. The observed false negatives indicate that these parameters should be calibrated according to the relative costs of missed anomalies, unnecessary investigations, and available review capacity. Lowering the operational threshold could increase sensitivity but could increase the false-positive rate, while raising it would reduce the number of reviews at the expense of additional missed anomalies. Taken together, the results address RQ1 by characterising the

framework's performance under controlled synthetic conditions, while highlighting metric-dependent trade-offs, uneven anomaly coverage, and the need for operational calibration.

5.2. Interpretability and Bayesian Uncertainty and RQ2

The statistical analysis layer addresses RQ2 by increasing transparency into the relationships underlying synthetic anomaly outcomes. Logistic regression achieved a ROC-AUC of 0.889. Its coefficients and odds ratios described the direction and magnitude of associations between the included predictors and planted anomaly status. Currency mismatch and recent bank-account change demonstrated the largest estimated associations. For continuous predictors, a one-training-standard-deviation increase in the paid-to-invoice ratio and invoice-to-PO ratio corresponded to odds ratios of 5.58 and 1.97, respectively, holding the other predictors fixed. In contrast, payment-term deviation resulted in an odds ratio below one, indicating a negative association with planted anomaly status after adjusting for other predictors. These findings elucidate the statistical relationships influencing anomaly status beyond a single classification score. However, the pronounced effects of bank-account change and currency mismatch are attributable to their deterministic construction within the synthetic dataset and should not be interpreted as real-world fraud-risk estimates. These coefficients describe a separately fitted model of planted anomaly status; they are not feature attributions for individual hybrid-score predictions.

Bayesian logistic regression improves interpretability by explicitly quantifying uncertainty in the estimated relationships between features. The Bayesian model used a reduced six-feature specification that excluded the bank-change and currency-mismatch indicators because of their deterministic relationship with planted anomaly status in the training data. Differences in predictive performance between the conventional and Bayesian models therefore reflect both their different predictor sets and their estimation procedures. Nevertheless, the estimated posterior results provide additional insight into the magnitude and uncertainty of individual associations. The paid-to-invoice ratio produced a posterior mean coefficient of 2.120, with an approximate 95% credible interval from 0.876 to 3.368 and a 99.9% posterior probability of a positive effect. The invoice-to-PO ratio also showed a consistently positive effect, while payment-term deviation yielded a posterior mean of -0.557 , with its credible interval entirely below zero. In contrast, the intervals for log payment amount, days-to-pay, and weekend payment included zero, indicating greater uncertainty about the direction of their associations.

Overall, the statistical findings support RQ2 regarding transparency and explicit uncertainty representation, rather than enhanced classification accuracy. Logistic regression offers an interpretable account of feature associations, while Bayesian modelling introduces approximate posterior distributions that quantify the precision of these estimates. This enables the framework to differentiate between the strength of an association and the certainty associated with it. Nevertheless, the Bayesian results

warrant caution, as mean-field ADVI yields an approximate posterior that may underrepresent dependencies or uncertainty among model parameters. Furthermore, the posterior predictions pertain to synthetic planted anomaly status within the specified model and should not be interpreted as calibrated probabilities of real-world fraud. Thus, RQ2 is supported to the extent that the statistical layer enhances interpretability and provides explicit uncertainty beyond what the operational alert mechanism alone offers.

5.3. Financial Decision Support and RQ3

The financial-analysis layer enhances PayGuard AI by providing monetary context to alerts generated by the detection framework, therefore extending its capabilities beyond conventional fraud classification. Bootstrap resampling of review-exposure values for the 66 held-out alerts yielded a mean aggregate exposure of approximately 302,564 SMU and a median of 286,918 SMU. The 95th percentile was 514,441 SMU, the 99th percentile was 634,443 SMU, and the conditional mean at or above the 95th percentile was approximately 587,257 SMU. These figures should not be interpreted as confirmed fraud losses, expected losses, or calibrated Value at Risk measures. Rather, they represent the financial magnitude associated with the review-exposure metric and illustrate how alert prioritisation can integrate both analytical risk evidence and monetary significance.

The Monte Carlo working-capital analysis addressed a distinct decision problem by assessing the potential financial outcomes of various supplier-payment timing strategies. Under the illustrative assumptions applied in the experiment, a 10-day Days Payable Outstanding (DPO) extension resulted in a mean net benefit of -25,171 SMU and a 37.2% probability of a positive outcome. The 20-day scenario yielded a mean benefit of 71,310 SMU with an 86.7% probability of a positive outcome, while the 30-day scenario produced the highest mean benefit at 166,905 SMU and a 97.8% probability of a positive result. The 10th percentile for the 30-day scenario also remained positive at approximately 71,272 SMU. These distributional results show the value of evaluating financial decisions under uncertainty, as opposed to relying solely on single deterministic estimates.

These findings support RQ3 at the proof-of-concept stage by demonstrating that transaction screening can be linked to subsequent financial considerations, including exposure, liquidity, and payment timing. The review-exposure analysis offers monetary context for flagged transactions, while bootstrap resampling quantifies uncertainty in aggregate exposure and Monte Carlo simulation captures uncertainty in alternative working-capital scenarios. Nevertheless, these outputs should be regarded as decision-relevant demonstrations rather than operational recommendations. The financial parameters were illustrative and annualised spend was calculated using synthetic ground-truth normal transactions. Practical implementation would require calibration based on organisation-specific financing costs, payment terms, discount structures, supplier relationships, historical investigation outcomes, and risk tolerance.

5.4. Practical Implications

PayGuard AI is intended to support investigation prioritisation with human review; its operational effectiveness has not yet been evaluated. The prototype produces the alert status, hybrid risk score, business-rule explanations, Isolation Forest anomaly indicators, interpretable statistical evidence, Bayesian uncertainty metrics, and estimated review exposure. These outputs are intended to help analysts interpret transaction priorities; their effects on analyst understanding and decision quality remain to be evaluated. Furthermore, the held-out alert rate of 3.93% demonstrates the framework's potential to substantially reduce the number of transactions requiring manual investigation, although the operational manageability of this alert volume would depend on organisational capacity. However, ultimate decisions regarding payment approval, rejection, or escalation remain subject to human judgment.

Additionally, PayGuard AI supports multiple financial functions. Accounts Payable teams benefit from efficient prioritisation of problematic transactions, while internal control and audit functions require transparent explanations and traceable evidence to substantiate alerts. Treasury professionals focus on liquidity, payment timing, and the financial implications of alternative strategies. PayGuard AI addresses these needs through a framework that integrates detection, interpretability, uncertainty, and financial-decision layers within a unified workflow. However, effective deployment necessitates integration with existing systems, governance protocols, and organisational investigation processes.

5.5. Limitations

The principal limitation of this study is its dependence on a fully synthetic transaction dataset. While the dataset was constructed to incorporate variation in suppliers, payments, currencies, timing, and purchase-order behaviour, it cannot capture the full complexity, temporal dynamics, organisational exceptions, and evolving fraudulent patterns found in authentic operational Accounts Payable data. Furthermore, the anomaly rate was fixed at 5%, and seven anomaly categories were deliberately constructed as a controlled experimental choice rather than reflecting empirically observed fraud frequencies. Certain indicators, including recent bank account changes and currency mismatches, were deterministically linked to the planted anomaly status. As a result, the reported performance metrics and feature relationships should be regarded as characteristics of the controlled experiment, not as estimates of expected behaviour in real corporate payment environments.

A second significant limitation concerns the calibration of the hybrid detector. The individual rule contributions, the 0.72/0.28 weighting between the rule-based and Isolation Forest components, and the 45-point alert threshold were selected as prototype settings rather than being optimised for investigation costs or operational outcomes. The limited detection of exact duplicates, near duplicates, and rapid-payment cases demonstrates the practical implications of these design decisions. Additionally,

the evaluation relied on a single dataset and a held-out partition, without external or temporal validation. Although the learned preprocessing and Isolation Forest model were fitted on the training partition, deterministic duplicate checks used the complete transaction batch. The experiment therefore does not represent strictly chronological screening using only information available before each payment. Future validation should use time-ordered evaluation and explicitly defined historical reference records. Consequently, the results reflect internal experimental behaviour and do not establish robustness across different organisations, time periods, or evolving transaction patterns.

Additional limitations concern the statistical analysis. The bank-change and currency-mismatch indicators were excluded from the Bayesian specification because each was deterministically linked to planted anomaly status when present in the synthetic training data. Consequently, the conventional and Bayesian logistic-regression models used different predictor sets, so their performance differences cannot be attributed solely to the estimation approach. Bayesian inference used mean-field ADVI, which may underrepresent posterior dependence and uncertainty. Completing all 30,000 optimisation iterations does not establish convergence or the accuracy of the variational approximation. The exceptionally large conventional logistic-regression odds ratios for bank-change and currency-mismatch indicators likewise reflect the synthetic construction rather than plausible real-world effect sizes.

The financial analyses presented are also subject to simplifying assumptions. Direct review exposure is calculated only when a measurable discrepancy, such as an overpayment or potential duplicate, is present. In contrast, alerts without an observable discrepancy rely on an illustrative scenario calculation using a prototype 0.25 factor. As a result, the resulting bootstrap distribution characterises variability in the defined review-exposure measure rather than actual fraud losses. Transaction-level resampling does not preserve potential dependence among payments from the same supplier or related duplicate-payment groups. Cluster-aware resampling should therefore be considered when evaluating operational data. Similarly, the Monte Carlo model employs illustrative probability distributions for financing costs, early-payment discounts, discount erosion, and supplier-friction costs. Annualised supplier spend is based on synthetic ground-truth normal transactions, representing an oracle assumption not available in typical operations prior to investigation. Therefore, the working-capital results should be interpreted as a demonstration of the simulation methodology rather than as a basis for prescribing a specific Days Payable Outstanding (DPO) strategy.

A further limitation is that the distributions of discount erosion and supplier-friction costs do not depend on the length of the DPO extension in the current simulation. The financing benefit increases linearly with the extension, while the cost distributions remain unchanged. Consequently, the higher expected net benefit for the 30-day scenario is partly built into the model structure and does not establish that a 30-day extension is optimal. An operational model should allow discount eligibility, discount

erosion and supplier-friction costs to vary with payment timing and contractual conditions, and should evaluate plausible dependence among these inputs.

5.6. Future Work

Future research should evaluate PayGuard AI using real, de-identified, and securely governed operational Accounts Payable data from multiple organisations. Temporal or out-of-time testing is particularly valuable for assessing changes in supplier behaviour and the framework's capacity to detect emerging anomaly patterns. Detection performance may be improved by empirically tuning near-duplicate weights and alert thresholds according to investigation costs and historical outcomes, rather than relying on fixed prototype values. The limitations observed in duplicate-detection performance underscore the necessity for more comprehensive duplicate features, including refinement of the existing fuzzy invoice-matching procedure, explicit historical reference records, and evaluation of alternative alert policies for duplicate-rule matches. Further investigation could incorporate temporal or graph-based supplier relationships and evaluate Bayesian models on datasets without deterministic links between individual indicators and anomaly labels, thereby facilitating a comparison between variational inference and full Markov Chain Monte Carlo (MCMC) posterior estimation.

Future research should also validate the financial and human decision-support components of the framework. Review-exposure assumptions may be calibrated using historical duplicate-payment recoveries, overpayments, and investigation outcomes. The working-capital simulation should integrate organisation-specific cost of capital, contractual payment terms, discount arrangements, and supplier-performance data. Evaluation involving Accounts Payable, audit, and treasury professionals would determine whether the explanations, uncertainty estimates, and financial outputs improve investigation efficiency and decision-making quality in practice. Such studies could assess usability, analyst trust, review workload, interpretation accuracy, and payment decision quality, thereby extending evaluation beyond model-level performance to address the organisational effectiveness of the complete decision-support framework.

6. Conclusion

This study developed and evaluated PayGuard AI, an explainable hybrid decision-support framework for corporate Accounts Payable fraud screening. The framework addresses limitations of relying solely on deterministic business rules or standalone machine-learning predictions by combining rule-based controls with Isolation Forest anomaly detection, interpretable statistical modelling, Bayesian uncertainty estimation, and financial decision-support analysis. Rather than treating transaction classification as the endpoint, PayGuard AI integrates detection with explanations, uncertainty estimates, review-exposure assessment, and working-capital simulation. This approach demonstrates how fraud screening can be embedded within a broader analytical workflow designed to support human financial investigation and decision-making.

Regarding RQ1, the framework demonstrated strong overall discrimination under controlled synthetic conditions. The hybrid achieved a ROC-AUC of 0.982, compared with 0.989 for the rule score alone, but yielded higher average precision (0.770 versus 0.680). At the selected threshold, precision was 84.8% and recall was 66.7%. Detection varied considerably across anomaly categories, with weaker identification of exact duplicates, near duplicates, and rapid-payment cases. These findings indicate metric-dependent trade-offs rather than uniform hybrid superiority and demonstrate the importance of anomaly-specific evaluation and operational calibration.

RQ2 is supported through the additional transparency and uncertainty information provided by the statistical layer. Logistic regression enables interpretable assessment of feature relationships, while Bayesian logistic regression supplements these estimates with approximate posterior uncertainty. The contribution of the Bayesian component, therefore, lies primarily in making uncertainty explicit rather than in improving classification accuracy. RQ3 is supported at the proof-of-concept level, as review-exposure analysis and financial simulation demonstrate how transaction-screening outputs can be extended to support subsequent investigation, exposure assessment, and working-capital decision-making.

The findings must nevertheless be interpreted in light of the experiment's controlled, synthetic nature and do not establish expected performance in operational Accounts Payable environments. Validation using real or securely de-identified transaction data, organisation-specific financial parameters, and evaluation with Accounts Payable, audit, and treasury professionals is required before practical deployment. Within these boundaries, PayGuard AI demonstrates the feasibility of integrating transaction detection, interpretability, uncertainty estimation, and financial analysis within a unified human-centred decision-support framework, providing a foundation for further research into transparent and financially informed corporate payment-risk systems.

Additional Diagnostic Results

The ADVI optimisation-loss trajectory is presented in Fig. 10. The loss trajectory shows the progress of variational optimisation. Completion of all 30,000 iterations does not establish that the parameter-convergence criterion was satisfied or that the variational approximation accurately represents posterior uncertainty. Approximate posterior anomaly probabilities and 95% credible intervals for the ten highest-risk held-out transactions are presented in Fig. 11.

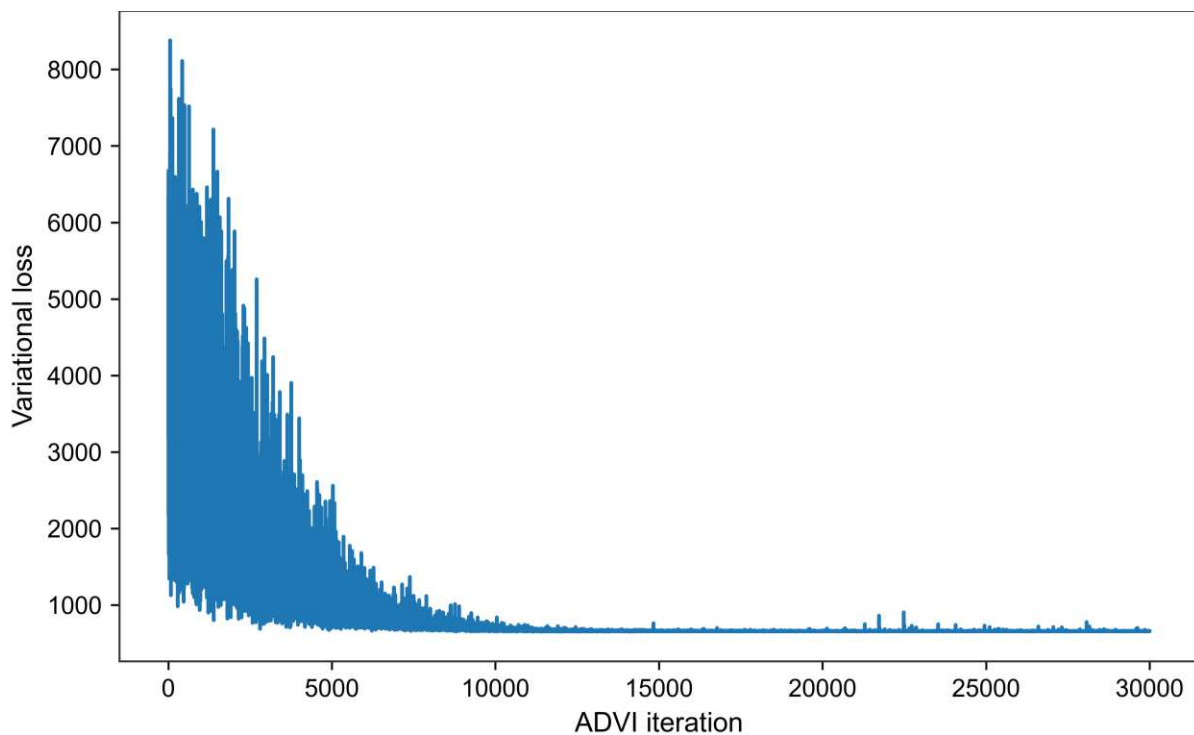


Fig. 10 ADVI optimisation-loss trajectory across 30,000 iterations

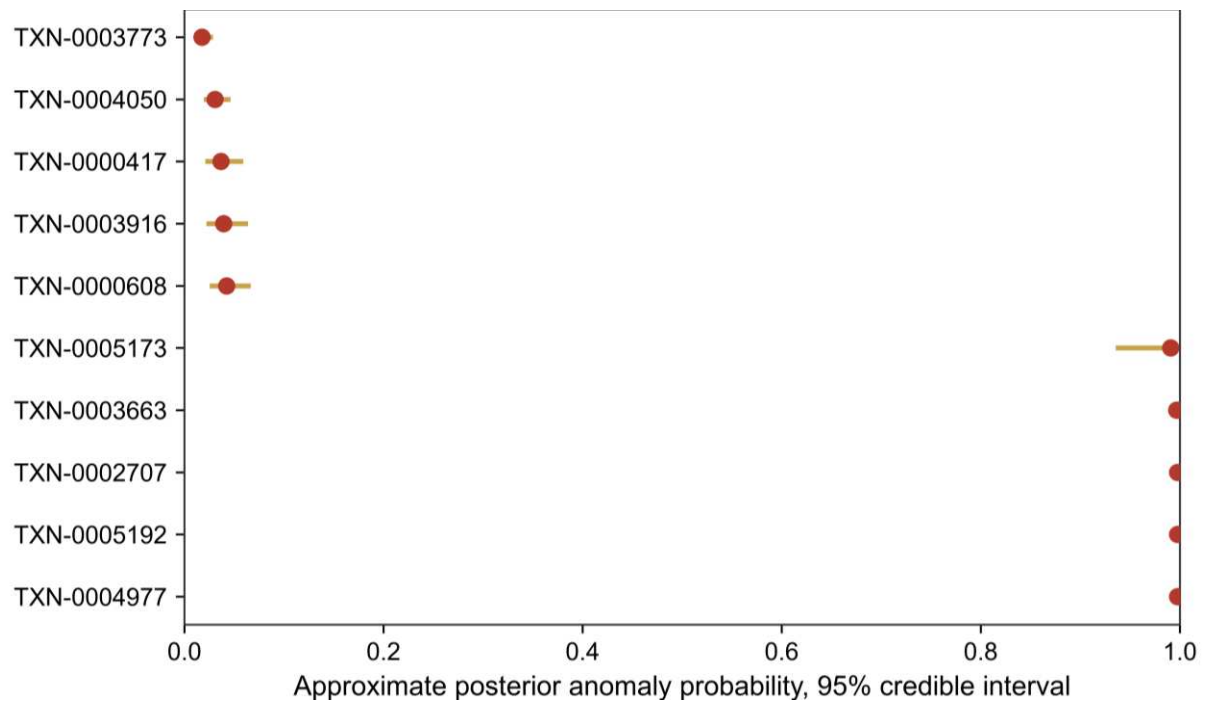


Fig. 11 Approximate posterior anomaly probabilities and 95% credible intervals for the ten highest-risk held-out transactions

Supplementary Materials

File S1: PayGuard AI reproducibility archive, containing the synthetic supplier-payment dataset, data dictionary, source code, dependency information, numerical results, and figure-generation material

Author Contributions

Conceptualization, T.D.; methodology, T.D.; software, T.D.; validation, T.D.; formal analysis, T.D.; investigation, T.D.; data curation, T.D.; writing - original draft preparation, T.D.; writing - review and editing, T.D.; visualization, T.D.; project administration, T.D. The author has read and approved the submitted manuscript and agrees to be accountable for all aspects of the work.

Funding

This research received no external funding.

Institutional Review Board Statement

Institutional review board approval was not required because the study used exclusively synthetic transaction data and did not involve human participants, personal data, or identifiable organisational records.

Informed Consent Statement

Not applicable.

Data Availability Statement

The synthetic dataset, data-generation code, analysis code, and materials supporting the findings of this study are publicly available in the PayGuard AI repository at <https://github.com/rockbjson/PayGuard-AI>. The repository includes documentation describing the dataset, analytical workflow, software dependencies, and reproduction of the reported results.

Acknowledgements

No artificial intelligence tools were used in the preparation of this manuscript.

Competing Interests

The author declares no competing interests.

References

- [1] Association for Financial Professionals, "2025 AFP Payments Fraud and Control Survey Report: Key Highlights," Rockville, MD, 2025.
<https://www.truist.com/content/dam/truist-bank/us/en/documents/info/cci/2025-afp-payments-fraud-control-survey-report-key-highlights.pdf>
- [2] Association of Certified Fraud Examiners, "Occupational Fraud 2024: A Report to the Nations," Association of Certified Fraud Examiners, Austin, Texas, 2024.
<https://www.acfe.com/en-us/-/media/files/acfe/pdfs/rttn/2024/2024-report-to-the-nations.pdf>
- [3] R. J. Bolton and D. J. Hand, "Statistical Fraud Detection: A Review," *Statistical Science*, vol. 17, no. 3, pp. 235-255, 2002. <https://doi.org/10.1214/ss/1042727940>
- [4] C. Phua, V. Lee, K. Smith and R. Gayler, "A Comprehensive Survey of Data Mining-based Fraud Detection Research," arXiv:1009.6119, 2010.
<https://doi.org/10.48550/arXiv.1009.6119>
- [5] E. W. T. Ngai, Y. Hu, Y. H. Wong, Y. Chen and X. Sun, "The Application of Data Mining Techniques in Financial Fraud Detection: A Classification Framework and an Academic Review of Literature," *Decision Support Systems*, vol. 50, no. 3, pp. 559–569, 2011. <https://doi.org/10.1016/j.dss.2010.08.006>
- [6] V. Chandola, A. Banerjee and V. Kumar, "Anomaly Detection: A Survey," *ACM Computing Surveys*, vol. 41, no. 3, pp. 1-58, 2009.
<https://doi.org/10.1145/1541880.1541882>
- [7] F. T. Liu, K. M. Ting and Z.-H. Zhou, "Isolation Forest," in *2008 Eighth IEEE International Conference on Data Mining*, Pisa, Italy, pp. 413-422, 2008.
<https://doi.org/10.1109/ICDM.2008.17>
- [8] A. Ali, S. A. Razak, S. H. Othman, T. A. E. Eisa, A. Al-Dhaqm, M. Nasser, T. Elhassan, H. Elshafie and A. Saif, "Financial Fraud Detection Based on Machine Learning: A Systematic Literature Review," *Applied Sciences*, vol. 12, no. 19, article 9637, 2022. <https://doi.org/10.3390/app12199637>
- [9] F. Doshi-Velez and B. Kim, "Towards a Rigorous Science of Interpretable Machine Learning," arXiv:1702.08608, 2017. <https://doi.org/10.48550/arXiv.1702.08608>
- [10] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti and D. Pedreschi, "A Survey of Methods for Explaining Black Box Models," *ACM Computing Surveys*, vol. 51, no. 5, article 93, pp. 1-42, 2018. <https://doi.org/10.1145/3236009>

- [11] Committee of Sponsoring Organizations of the Treadway Commission, "Internal Control—Integrated Framework: Executive Summary," Committee of Sponsoring Organizations of the Treadway Commission, 2013.
- [12] P. Ravisankar, V. Ravi, G. Raghava Rao and I. Bose, "Detection of Financial Statement Fraud and Feature Selection Using Data Mining Techniques," *Decision Support Systems*, vol. 50, no. 2, pp. 491-500, 2011. <https://doi.org/10.1016/j.dss.2010.11.006>
- [13] J. L. Leevy, J. Hancock, T. M. Khoshgoftaar and A. Abdollah Zadeh, "Investigating the effectiveness of one-class and binary classification for fraud detection," *Journal of Big Data*, vol. 10, p. 157, 2023. <https://doi.org/10.1186/s40537-023-00825-1>
- [14] H. He and E. A. Garcia, "Learning from Imbalanced Data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263-1284, 2009. <https://doi.org/10.1109/TKDE.2008.239>
- [15] T. Fawcett, "An Introduction to ROC Analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861-874, 2006. <https://doi.org/10.1016/j.patrec.2005.10.010>
- [16] W. Zhou and G. Kapoor, "Detecting Evolutionary Financial Statement Fraud," *Decision Support Systems*, vol. 50, no. 3, pp. 570-575, 2011. <https://doi.org/10.1016/j.dss.2010.08.007>
- [17] D. W. Hosmer, S. Lemeshow and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed., John Wiley & Sons, 2013. <https://doi.org/10.1002/9781118548387>
- [18] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5-32, 2001. <https://doi.org/10.1023/A:1010933404324>
- [19] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila and F. Herrera, "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI," *Information Fusion*, vol. 58, pp. 82-115, 2020. <https://doi.org/10.1016/j.inffus.2019.12.012>
- [20] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA: ACM, 2016, pp. 785–794. <https://doi.org/10.1145/2939672.2939785>
- [21] S. Bhattacharya, D. Xu and K. Kumar, "An ANN-based Auditor Decision Support System Using Benford's Law," *Decision Support Systems*, vol. 50, no. 3, pp. 576-584, 2011. <https://doi.org/10.1016/j.dss.2010.08.011>

- [22] I. Goodfellow, Y. Bengio and A. Courville, *Deep Learning*, The MIT Press, 2016.
<https://www.deeplearningbook.org>
- [23] M. M. Breunig, H.-P. Kriegel, R. T. Ng and J. Sander, "LOF: Identifying Density-Based Local Outliers," in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, Dallas, TX, USA: ACM, 2000, pp. 93–104.
<https://doi.org/10.1145/342009.335388>
- [24] J. Černevičienė and A. Kabašinskas, "Explainable artificial intelligence (XAI) in finance: a systematic literature review," *Artificial Intelligence Review*, vol. 57, p. 216, 2024. <https://doi.org/10.1007/s10462-024-10854-8>
- [25] C. Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206-215, 2019. <https://doi.org/10.1038/s42256-019-0048-x>
- [26] R. Sabharwal, S. J. Miah, S. F. Wamba and P. Cook, "Extending application of explainable artificial intelligence for managers in financial organizations," *Annals of Operations Research*, vol. 354, pp. 309–339, 2025.
<https://doi.org/10.1007/s10479-024-05825-9>
- [27] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*, Cambridge: The MIT Press, 2012.
- [28] J. C. Hull, *Risk Management and Financial Institutions*, 5th ed., Wiley, 2018.
- [29] P. Glasserman, *Monte Carlo Methods in Financial Engineering*, vol. 53, New York, NY: Springer, 2003. <https://doi.org/10.1007/978-0-387-21617-1>
- [30] P. Jorion, *Value at Risk: The New Benchmark for Managing Financial Risk*, 3rd ed., New York: McGraw-Hill, 2007.
- [31] R. T. Rockafellar and S. Uryasev, "Optimization of Conditional Value-at-Risk," *The Journal of Risk*, vol. 2, no. 3, pp. 21-41, 2000.
<https://doi.org/10.21314/JOR.2000.038>
- [32] M. Deloof, "Does Working Capital Management Affect Profitability of Belgian Firms?," *Journal of Business Finance & Accounting*, vol. 30, no. 3-4, pp. 573-588, 2003. <https://doi.org/10.1111/1468-5957.00008>
- [33] R. I. T. Jensen, J. Ferwerda, K. S. Jørgensen, E. R. Jensen, M. Borg, M. P. Krogh, J. B. Jensen and A. Iosifidis, "A synthetic data set to benchmark anti-money laundering methods," *Scientific Data*, vol. 10, no. 1, p. 661, 2023.
<https://doi.org/10.1038/s41597-023-02569-2>