

## Article

# Enhancing Enterprise Risk Management and Internal Audit Practices by Applying Machine Learning Models

Reneta Duhova<sup>1,\*</sup>, Angel Duhov<sup>1,\*</sup> , Petia Georgieva<sup>2</sup>  and Milena Lazarova<sup>1</sup> 

<sup>1</sup> Department Computer Systems, Faculty of Computer Systems and Technologies, Technical University of Sofia, 1000 Sofia, Bulgaria; reneta.duhova@tu-sofia.bg (R.D.); milaz@tu-sofia.bg (M.L.)

<sup>2</sup> Department of Electronics Telecommunications and Informatics, Institute of Electronics and Informatics Engineering of Aveiro (IEETA), University of Aveiro, 3800-193 Aveiro, Portugal; petia@ua.pt

\* Correspondence: aduhov@tu-sofia.bg

## Abstract

Organizations are currently in a stage where the volume of financial transactions and data is constantly growing. The same goes for risks associated with the use of data for risk management and strategic decision-making. The likelihood of transactional errors generally increases with data volume and process complexity, while fraud, although less frequent, may have more severe financial, compliance, and reputational consequences for organizations. Continuous auditing practices and well-established enterprise risk management (ERM) processes, combined with AI-driven pattern recognition, trend analysis and segmentation, can enhance timely detection and proper investigation of suspicious transactions. In areas with large volumes of transactions, the audit sampling process may be a lengthy process and pose a detection risk. Using machine learning (ML) models to support critical business processes could prove effective in managing enterprise risk overall. The current study offers new perspectives on managing risk and assurance with ML model output for flagging possible risky transactions within ERP (SAP) systems data. The study population consists of 69,158 finalized billing records extracted from the SAP production environment of a private sector organization, which covers a six-month operational period. The dataset was divided into an 80/20 train–test split, yielding 55,326 training and 13,832 test instances across six classification categories. The study examines the ML methods' outcomes from billing datasets and their applicability in enhancing audit, assurance, and ERM processes by evaluating output data results from two supervised classification algorithms—multinomial logistic regression (SoftMax regression) and XGBoost—against various criteria generally accepted as risky in audit engagements. Model performance was assessed using accuracy, precision, recall, F1-score, ROC-AUC, and average precision (AP) from precision–recall curves. The results confirm that XGBoost achieves 99% overall accuracy with a macro F1-score of 0.965, outperforming logistic regression (macro F1 = 0.863), and that ML output allows early investigation and follow-up procedures to minimize the risk of fraud and errors and optimize risk management activities, thus strengthening internal control frameworks.



Academic Editor: Paolo Giudici

Received: 31 March 2026

Revised: 25 April 2026

Accepted: 28 April 2026

Published: 6 May 2026

**Copyright:** © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

**Keywords:** artificial intelligence; machine learning models; enterprise risk management; continuous auditing; SoftMax regression; XGBoost; enterprise resource planning

## 1. Introduction

With the constantly growing volume of financial transactions, the digital transformation of business units and the ongoing expansion of the development and usability of

artificial intelligence (AI) techniques in different business areas, companies can benefit from implementing machine learning (ML) models in transaction monitoring and early fraud detection. Thus, they can optimize enterprise risk management processes and strengthen internal control practices. The same models can be explored as well by internal audit departments by implementing continuous audit practices in the identified high-risk areas of the businesses.

Traditional auditing practices may fall short in detecting fraudulent accounting activities since most managers recognize the limitations of audits (Jofre and Gerlach 2018), highlighting the necessity for additional dynamic and comprehensive analytical approaches to accurately identifying accounting fraud at an early stage.

Current threats in enterprise risk management (ERM) as a governance process include cyberattacks, data breaches, fraud, supply chain disruptions, regulatory framework changes and uncertainty arising from rapid technological evolution. Emerging risks from artificial intelligence and machine learning—such as novel models, ethical, and operational risks—require vigilant governance and adaptive controls to maintain resilience in dynamic environments (Di Palma et al. 2025).

The integration of artificial intelligence, big data, and cloud computing is consolidating a paradigm shift in organizational processes and how current professionals conduct their activities (Shapovalova et al. 2023).

AI has a set of capabilities, including real-time information processing, trend identification and task automation (Langley 2011). AI-powered systems have the processing capability of analyzing large sets of data to support in identifying anomalies such as unusual transaction patterns or deviations from normal customer behavior. ML algorithms can flag potentially fraudulent activities in near real-time, allowing for swift investigation and intervention. Additionally, AI could boost audit practices by automating data analysis and reducing the detection risk in audit engagements.

By integrating ML models into internal audit workflows, organizations can minimize detection risk by automating transaction monitoring (Knoblauch and Großmann 2023); in addition to that, risk management processes can be enhanced, allowing proactive actions by business units. The paper provides an overview of the current state of the art in auditing, sampling and ERM fundamentals and describes machine learning models used for fieldwork. Then, it presents a comparison of the different ML algorithms' output against audit-relevant risk criteria derived from ERP system data. The risk criteria capture patterned transaction amounts, manual activity outside standard working hours, deviations from customer-specific shipping conditions, edge-of-month billing concentration and missing VAT information after the applicable threshold.

The novelty of this study does not lie in proposing a new machine learning algorithm, as multinomial logistic regression and XGBoost are already established in fraud detection research. Rather, the contribution lies in operationalizing these methods within an ERP-native SAP billing environment by transforming raw transactional attributes into audit-engineered variables and mapping them to clearly defined audit risk classes.

Despite the increasing application of machine learning techniques in fraud detection and transaction analytics, several important limitations remain in the current literature. First, many studies focus primarily on predictive performance without explicitly linking model outputs to ERM processes and internal audit workflows. Second, existing research often relies on generic datasets or simulated environments, while studies based on ERP-native transactional data—particularly within real organizational contexts—remain relatively limited. Third, there is a lack of structured approaches that integrate audit-relevant domain knowledge, such as control indicators and risk-based variables, into the design of machine learning models.

In response to these limitations, this study makes several contributions. (1) We define an audit-oriented anomaly taxonomy that operationalizes common continuous-auditing red flags into mutually exclusive classes, enabling systematic population-level monitoring rather than ad-hoc sampling. (2) We construct a set of audit-engineered variables derived from ERP billing data that explicitly reflect business-process risk characteristics and support transparent interpretation of flagged cases. (3) We provide a comparative evaluation between an interpretable baseline (multinomial logistic regression) and a non-linear ensemble (XGBoost) under the same feature set, emphasizing alert quality under class imbalance through precision–recall analysis. (4) We discuss how the resulting outputs can be embedded into assurance and ERM routines, supporting earlier investigation, workload-aware triage, and governance-oriented documentation.

The remainder of the paper is organized as follows. Section 2 presents the current state and reviews related work on enterprise resource planning, continuous auditing, enterprise risk management and utilization of machine learning in these fields. Section 3 presents the audit case study researched in the paper. Sections 4 and 5 discuss the proposed models' design and implementation based on multinomial logistic regression and XGBoost, respectively. Section 6 contains a discussion on models' comparison, and Section 7 describes possible future work areas and concludes the paper.

## 2. Related Work

### 2.1. Enterprise Resource Planning

Enterprise Resource Planning (ERP) systems are integrated enterprise information systems that support the execution, coordination, and monitoring of core organizational processes, including procure-to-pay, order-to-cash, record-to-report, manufacturing, logistics, and human resources (Grabski et al. 2011). Their primary value proposition lies in standardizing process logic and consolidating transactional and master data within a unified architecture. This integration reduces informational silos, improves cross-functional visibility, and strengthens data consistency across organizational units, which is essential for timely managerial decision-making and effective governance. Among the pioneering companies in this field, SAP (Systems, Applications, and Products (in Data Processing)) played a significant role in the development of standardized business software and continues to provide industry-leading ERP solutions (SAP SE 2024).

In contrast to fragmented application landscapes, where departments maintain separate operational databases and heterogeneous process definitions, ERP environments enable end-to-end traceability through harmonized data structures, shared master data, and consistent posting rules. Beyond operational efficiency, this integrated design has direct relevance for risk management and assurance functions. ERP systems generate detailed, time-stamped transactional records and process-relevant attributes that can serve as a foundation for continuous monitoring, control testing, and anomaly detection. Consequently, ERP data provides a suitable basis for machine learning approaches that aim to identify deviations from expected process behavior and flag potentially risky transactions for further review.

### 2.2. Continuous Auditing

With the increasing complexity of companies' structures and the growing number of policies and procedures, transactions and IT systems in use, the traditional audit approach is falling behind in practical results and timely addressed issues, and its effectiveness is decreasing (Alles et al. 2006). The need for real-time assurance and prompt follow-ups on risk areas prepares the ground for a continuous auditing approach.

The continuous auditing (CA) methodology possesses the capacity to improve both the efficiency and accuracy of audit practices by supporting real-time assurance processes. It is generally accepted that the traditional audit approach struggles to provide real-time assurance over business processes and procedures because of its retrospective evidence. In the traditional approach, many audit procedures are carried out manually, making them time-consuming. Because of these constraints, traditional audits are often scheduled and executed on an annual basis or by a pre-defined audit plan. This limitation may result in a lack of valuable recommendations to stakeholders. However, those limitations can be overcome through the adoption of technological advancements and automation. A brief comparison of traditional and continuous audit is presented in Table 1.

**Table 1.** Traditional and continuous audit comparison.

| Features   | Traditional Audit                  | Continuous Audit             |
|------------|------------------------------------|------------------------------|
| Frequency  | Periodic, pre-defined              | Close to real time           |
| Approach   | Reactive                           | Proactive                    |
| Procedures | Manual, semi-automated             | Automated, semi-automated    |
| Audit role | Assurance, optimization            | Risk reduction, optimization |
| Coverage   | Sample                             | Entire population            |
| Reporting  | Pre-defined, according to the plan | Continuous, ad-hoc           |

There is an option to reduce the manual nature of the audit process and also to increase efficiencies by relying on technology (Zeng and Li 2025). Continuous audit frameworks can prove to be effective in enhancing audit procedures and providing greater coverage and assurance over identified risk areas. On the other hand, continuous auditing not only streamlines the audit process but also provides valuable, timely alerts and insights to internal stakeholders (Aladwey and El Sayad 2024).

By implementing continuous auditing practices and procedures in organizations utilizing ERP systems, internal auditors can boost the probability of detecting anomalies, assess control effectiveness, and provide timely, data-driven insights that strengthen overall governance and risk management. Continuous auditing differs in various instances from continuous monitoring (CM). In many cases, they have, ultimately, the same end goal—strengthening the internal control system. However, the main difference is in the ownership of the process. CA is applied by only the internal auditors (Davidson et al. 2013), whereas CM is established at the management level.

Continuous monitoring is a structured management process that monitors whether internal controls are operating effectively as designed, and managers are advised to interpret high scores with caution and to support qualitative evaluations (Ahmad et al. 2025). It is executed by operating management, which means continuous monitoring lies in the first and second lines of defense in the entity.

Continuous auditing is a process defined and executed by internal audit, and it is designed to enable the internal auditor to report on subject matter within a much shorter time frame than under the traditional retrospective approach.

### 2.2.1. Benefits and Challenges

The main benefits of the continuous audit practices are the proactive and real-time monitoring of audited business processes. The higher probability of capturing fraud and/or mistakes in financial statements and business reporting benefits not only the audit department but also the company's risk management. Organizations benefit from technology-driven automated audit routines, providing auditors with rapid access to relevant data and facilitating the identification of anomalies as they occur (Brown et al. 2007).

On the other hand, there are several challenges that need to be addressed while implementing continuous audit processes within an organization. Implementation challenges are related to organizations' complexity, potential concerns around internal auditors' independence (Polizzi and Scannella 2023) and the issue of considering a continuous auditing system as a real-time error correction mechanism. The successful implementation of continuous auditing requires upfront investment in technology development and integration, systems compatibility, and the development of robust frameworks. On the other hand, bias and ethics are critical considerations in AI systems applied for auditing, as flagged results may disproportionately affect certain transaction types or areas (Murikah et al. 2024). ML/AI systems may be vulnerable to inappropriate human intervention if adequate governance controls are not established. This may include unauthorized changes to input data, feature-engineering logic, model parameters, decision thresholds, classification rules, or reported outputs. Organizations can mitigate this risk through role-based access controls, segregation of duties, version control, audit trails, data lineage documentation, and periodic independent validation. These measures support model integrity, reproducibility, and auditability.

#### 2.2.2. EU AI Act for AI and ML Regulation

In recent years, the EU has played a leading role in digital and AI governance through the GDPR, which strengthened personal data protection and accountability, and the EU AI Act, which introduces a harmonized, risk-based framework for AI systems across EU Member States. European data protection law is recognized as a standard for data protection processes and policies. The EU approach to AI regulation is widely expected to shape how countries will approach the challenges of AI governance (Tarafter and Vadlamani 2024; Bueno and Canaan 2024).

The EU AI Act classifies AI systems by risk tiers. While transaction-monitoring ML models for fraud detection are exempted from the high-risk Annex III list (Recital 58), they still require robust governance under limited-risk obligations, including transparency, data quality, and human oversight. High-risk AI (e.g., credit scoring) must undergo thorough evaluation of build cycles, data input management, and output reproducibility. For exempted models like these, aligning with IT control best practices remains essential: robust data governance, versioned documentation, and auditability ensure integrity. Internal/external audits must recreate model outputs, mirroring internal control frameworks.

This regulation's demands align closely with IT control best practices: ML systems must employ robust data-governance controls and clear, versioned documentation of model logic to satisfy transparency requirements. Audit bodies (internal and external) must be able to recreate ML models' output. Moreover, detailed documentation of model design, training data, and risk mitigation strategies supports auditability, mirroring requirements in internal control frameworks.

#### 2.3. Enterprise Risk Management

Enterprise Risk Management (ERM) refers to an integrated, organization-wide approach to identifying, assessing, managing, and monitoring risks that could affect an entity's achievement of objectives or threaten its value (Akbaş 2024). ERM frameworks provide principles and guidelines for establishing a structured and systematic risk management process that is embedded within strategic and operational activities across the enterprise (Ispas et al. 2023). As organizations face volatile markets, complex regulations, such as changes in financial reporting, emerging digital technologies regulations, data transfer limitations, and technology-driven disruptions, ERM has evolved from a compliance-driven

activity to a strategic imperative, guided by real-time analytics and fostering a risk-aware culture (Bakos and Dumitraşcu 2021).

In its recent form, ERM is characterized by advanced integration of digital tools, including artificial intelligence and machine learning, which enable dynamic risk assessment, automated monitoring, and predictive modeling for emerging threats (Kumar 2022). These technologies facilitate deeper risk insights, improve scenario planning, and support agile responses to rapidly changing environments (Le et al. 2024). Notably, ERM now plays a vital role in supporting corporate governance and stakeholder confidence, particularly as reporting standards demand greater transparency and proactive risk disclosure (Xiao et al. 2023).

ERM brings several benefits, including enhanced decision-making, reduced operational losses, improved compliance, and the potential for competitive advantage through risk-informed strategy (Gurtu and Johnny 2021). However, the implementation of enterprise-wide risk management is not without challenges. Firms often encounter difficulties with risk data quality, organizational silos, model validation, and maintaining alignment between ERM processes and overall business objectives (Locurcio et al. 2021).

Moreover, while digitalization enhances ERM's effectiveness, it also introduces new concerns such as cyber risk, data privacy, and the complexity of managing AI-driven risks across global operations. As a result, organizations must invest in ongoing staff training, strong governance structures, and technology integration to fully realize ERM's potential while managing its limitations (Kumar 2022).

While ERM frameworks may vary in structure and terminology, they are generalized into six components.

#### 2.3.1. Risk Governance and Oversight

They establish clear roles and responsibilities for managing risks throughout their life cycle within the organization, often through a dedicated risk management committee, containing professionals with diverse backgrounds representing involved parties.

#### 2.3.2. Risk Identification

It involves an ongoing process of identifying potential risks that could affect organizational goals and objectives, using different techniques (Bakos and Dumitraşcu 2021). Such activities are performed on a recurring basis to ensure timely capture of possible risks.

#### 2.3.3. Risk Assessment

It evaluates the likelihood and magnitude of potential impact for each of the identified risks, using qualitative and quantitative techniques. Risk assessment can be done by different methods, including qualitative, quantitative or semi-quantitative.

#### 2.3.4. Risk Appetite and Tolerance

They determine the organization's risk appetite—the amount of risk it is willing to accept for executing its core business functions—and risk tolerance levels for different types of risks. Organizations within the same sector can have different risk appetites based on their maturity, targets and organizational structure.

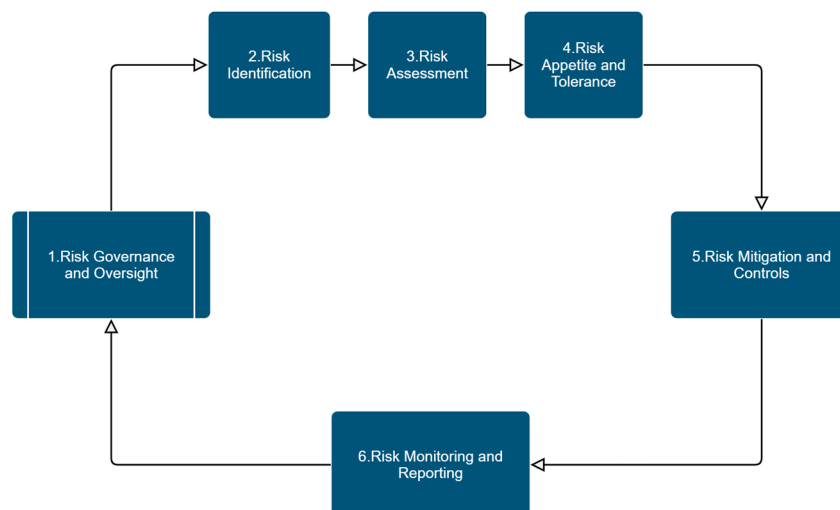
#### 2.3.5. Risk Mitigation and Controls

**Risk Mitigation and Controls:** Develop and implement strategies, defining the proper sequence of actions to manage identified risks effectively, including preventive measures such as risk transfer, risk acceptance, risk mitigation and risk avoidance.

### 2.3.6. Risk Monitoring and Reporting

They establish mechanisms for continuous monitoring of risk factors and performance indicators and regular reporting of risk information to stakeholders, allowing timely and informed decision-making on corrective actions.

All the components described are summarized and shown in Figure 1.



**Figure 1.** Enterprise risk management's main components (process flow).

### 2.4. Enterprise Risk Management Frameworks

Enterprise risk management is commonly structured through established frameworks that provide organizations with principles, processes, and governance mechanisms for identifying, assessing, monitoring, and responding to risks, where two of the most recognized frameworks are the COSO Enterprise Risk Management framework and ISO 31000.

The COSO ERM framework emphasizes the integration of risk management with strategy, performance, governance, review, and information flows (COSO 2018). Its updated version, Enterprise Risk Management—Integrating with Strategy and Performance, highlights the importance of considering risk both in strategy-setting and in driving organizational performance. This perspective is directly relevant to the present study, as ML-based transaction monitoring can support risk identification, control monitoring, and timely reporting within ERM processes.

ISO 31000:2018 provides general guidelines for risk management and outlines a structured approach to identifying, analyzing, evaluating, treating, monitoring, and communicating risks across an organization (ISO 2018). In contrast to a narrow compliance-oriented view, ISO 31000 supports the integration of risk management into governance, decision-making, and organizational processes. From this perspective, machine learning outputs should not be interpreted as isolated technical predictions, but as risk signals that can support risk assessment, prioritization, monitoring, and follow-up actions.

### 2.5. ML Models for Enterprise Risk Management

There are various related studies in this area, each one of them bringing different findings and observations into using ML models for fraud detection or analytics of big data. Different studies enforce deep learning methods or log analysis for capturing unusual data patterns or inconsistencies in generally accepted data structures.

As suggested by one relevant study, evaluate supervised (e.g., XGBoost, RF) and unsupervised ML algorithms on real general ledger data to detect financial anomalies, achieving high F1-scores on journal entry lengths via vectorization (Bakumenko and Edman

2022). The study highlights XGBoost's superiority for imbalanced data, reducing manual review. Practical results suggest that organizations can deploy this hybrid approach for proactive fraud detection, but should carefully address challenges around data quality. Data integrity input, sanitization and output need to be constantly monitored to avoid bias in the output or data privacy violations.

Another relevant research in the field presents fraud detection capabilities of SAP implementation by capturing audit logs and trails for changes in master data of vendor records (Best et al. 2009). The approach includes threat monitoring—routine scanning of security audit logs for critical transaction patterns—and automated profiling of user activity to establish suspicious behavior. Their method uses forensic techniques to validate anomalies and has been tested in a real SAP case, illustrating early identification and documentation of actual fraud events. The authors show that continuous monitoring with audit trails provides a strong basis for ongoing assurance; however, limitations on so-called super user activity, who can erase or modify audit logs, need to be taken into consideration.

A different study, which overviews AI models of the Big Four audit and assurance firms, concludes that adopting machine learning, and specifically the Random Forest algorithm, can dramatically enhance the accuracy and efficiency of high-risk identification in enterprise financial audits, outperforming SVM and KNN models (Yuan et al. 2025). The analysis also finds that audit frequency, historical compliance violations, employee workload, and client satisfaction are the most influential predictors of audit risks. Key components contributing to models' efficiency are feature engineering and data processing. Despite these strengths, the study acknowledges its reliance on structured, internal audit variables and the omission of text analytics or external market signals—highlighting a gap in capturing nuanced risk factors and emerging threats.

While prior research has demonstrated the effectiveness of machine learning models for anomaly detection and fraud identification, the integration of these techniques within enterprise risk management and internal audit contexts remains underdeveloped. In particular, limited attention has been given to aligning model design with audit-specific risk indicators, control frameworks, and ERP-based transactional environments. The present study addresses these gaps by combining ERP-derived audit variables, supervised learning models, and an evaluation framework that explicitly considers their applicability in continuous auditing and ERM processes.

### 3. ML-Based Continuous Auditing

#### 3.1. Machine Learning

According to the generally accepted definition, machine learning refers to the study of giving computers the ability to learn without being explicitly programmed (Samuel 1959). Building on the field of artificial intelligence, machine learning researchers gradually shifted their focus from serving as a tool for AI to solving practical problems using models from probability theory and statistics (Li et al. 2023). Several types of problems for ML models include classification, regression and clustering. In the context of auditing, assurance and risk management, ML offers powerful capabilities to automate anomaly detection, flag suspicious transactions and reduce detection risk. Supervised learning, the primary approach examined in this article, involves training a model on a labeled dataset, where the input features are paired with known outcomes (or classes).

The practical part of the study makes a comparison between two machine learning algorithms, which are considered applicable for classification tasks with multiple classes. Their mathematical equations and general description are outlined below.

### 3.1.1. Logistic Regression

Logistic regression, a widely used supervised learning algorithm, is suitable for binary classification tasks because of its simplicity, interpretability, and solid performance (Brandão and Georgieva 2020). The multi-class logistic regression model in this project uses the SoftMax function, also referred to as SoftMax regression, which extends the traditional logistic regression to handle classification with more than two classes. The mathematical equation on which the model is relying is shown below:

$$(y = k \ x) = \frac{\exp(w_k^T x + b_k)}{\sum_{j=1}^K \exp(w_j^T x + b_j)}, \quad (1)$$

where  $x$  is the feature vector, i.e., the preprocessed input data,  $w_k$  is the weight vector for class  $k$ ,  $b_k$  is the bias term for class  $k$ ,  $K$  is the total number of classes.

### 3.1.2. XGBoost Classification

The XGBoost (Extreme Gradient Boosting) classifier is a widely used ensemble learning model that builds decision trees sequentially, optimizing for prediction accuracy by minimizing loss with each added tree. XGBoost is renowned for its speed, scalability, and high predictive power, making it particularly suitable for large, complex datasets in domains such as healthcare diagnosis, fraud detection, and risk assessment. In medical applications, for example, XGBoost has demonstrated excellent performance in early-stage disease prediction, offering high accuracy, sensitivity, precision, and AUC, and has become integral in predictive modeling for population-scale studies (Gündoğdu 2023).

$$\hat{y}_i = \sum_{t=1}^T f_t(x_i), \quad f_t \in \mathcal{F} \quad (2)$$

$$L = \sum_{i=1}^n l(y_i \hat{y}_i) + \sum_{t=1}^T \Omega(f_t) \quad (3)$$

$$\Omega(f) = \gamma M + \frac{\lambda}{2} \sum_{j=1}^M \omega_j^2 \quad (4)$$

where:

- $x_i$  is the input feature vector,  $y_i$  is the true target,  $n$  is the number of samples.
- $\hat{y}_i$  is the model prediction (score/logit, later mapped to a class probability in classification).
- $T$  is the number of trees (boosting rounds).
- $f_t$  is the  $t$ -th regression tree,  $\mathcal{F}$  is the set of all regression trees.  $M$  is the number of leaves in a tree,  $w_j$  is the score assigned to leaf  $j$ .
- $l(\cdot)$  is the loss function (for example, logistic loss or softmax cross-entropy).
- $\Omega(\cdot)$  is the regularization term, with  $\gamma$  penalizing tree complexity and  $\lambda$  controlling L2 regularization on leaf weights.

### 3.2. Auditability and Integrity of ML Models

Security of ML models and integration with related systems is a key element of maintaining a strong internal control system. Having confidence in security allows greater reliance on the ML model output data. For ML models to provide reasonable data that is later used for follow-up activities and decision-making, there must be assurance of the model's integrity, which means that the ML model needs to be audited by skilled professionals. Audits could be conducted either by an internal audit department or by a

third-party company, which is considered useful for higher-risk applications that support critical business operations.

Although expensive, a third-party audit results in a higher level of assurance among various system stakeholders (Knowles and Richards 2021). However, internal audit can provide additional benefits for overall processes, as it encourages the creation of documentation and supports internal risk analysis.

ML and AI application models currently rely heavily on training data (ISACA 2022). Therefore, when performing an audit, evaluating the data sources is the critical first step of ML auditing. At that stage, the focus needs to be on data compliance and governance. It is critical for the auditors to assess whether the data sources are compliant with the applicable laws and regulations.

### Auditor's Role in Validating ML Models

The use of machine learning outputs in assurance contexts should be consistent with core auditing principles outlined by International Auditing Standards (ISA) related to risk assessment (IAASB 2019), professional judgment (IAASB 2009a, 2009b) and the requirement to obtain sufficient appropriate audit evidence (IAASB 2009c). In this study, ML-based flagging is positioned as a decision-support mechanism that helps prioritize transactions for review within continuous auditing and ERM monitoring, rather than as a replacement for auditor judgment.

Accordingly, auditors (or an independent assurance function) have a validation and governance role prior to deployment and during ongoing use. Key responsibilities include verifying data provenance and the integrity of ERP extracts, confirming that engineered indicators and class definitions reflect auditable control expectations, validating performance using imbalance-aware metrics, and maintaining an evidence trail that supports reproducibility, documentation, and periodic monitoring for drift. This framing supports controlled adoption by linking model outputs to transparent validation steps and human oversight.

### 3.3. Challenges and Limitations of ML Models

There are several challenges and limitations of ML models that need to be considered in the long term when using the model.

- **Quality of the data:** In production datasets, and given the fact that there are a lot of customized features and functionalities, verifying the reliability, integrity and completeness of the data is essential for all users of the ML models. Low data quality, like missing fields, inconsistencies and errors, can undermine model performance and reliability (Shin et al. 2013). Their prevalence depends on organizational factors such as ERP configuration, data governance maturity, process standardization, and internal control effectiveness. Therefore, these issues are discussed as commonly reported implementation risks rather than as measured frequencies in the present dataset.
- **Integration:** Integrating machine learning models with existing ERP systems and workflows requires careful planning and implementation (Țîrcovnicu and Hătegan 2023). Compatibility challenges, effective data synchronization, and ongoing efforts to maintain system stability during machine learning model integration are key factors. In addition, there should be ongoing monitoring of integration protocols, version control across backend platforms, and the resolution of dependencies to ensure seamless communication and sustained performance between ML models and core enterprise systems. This ongoing monitoring should be designed so that automated systems can detect flag deviations in integration protocols, version mismatches, and dependency conflicts, but the interpretation of findings and the authorization of corrective actions

remain the responsibility of qualified IT and audit professionals. Full delegation of this oversight to automated processes alone is not recommended without appropriate governance controls and defined escalation procedures in place.

- **ML model security:** Data poisoning, adversarial attacks and data privacy risks are among the most critical IT security concerns for machine learning models. Data poisoning involves injecting malicious or misleading information into training sets, which can compromise model accuracy. Adversarial attacks exploit specially crafted inputs to cause models to misclassify or behave unpredictably, undermining reliability. Data privacy risks emerge when sensitive information used in model training is leaked or inferable from outputs, endangering organizational and user confidentiality.

As all those challenges need to be addressed, there are several limitations of ML models that must be considered when put to use to maintain audit objectivity and perform the right follow-up actions on the model results.

- **Adaptability:** Machine learning models trained in historical data may not be as precise when, for example, more classes are introduced. Continuous model monitoring and validation of the provided results are necessary to maintain effectiveness with a more complex set of rules for consideration.
- **Disbalance in classes:** Risky transactions are much fewer in comparison to legitimate ones, which may result in class imbalance in the dataset. This imbalance can result in models that are biased toward the majority class and have reduced sensitivity to detect fraud instances (Berisha et al. 2022).

### 3.4. Continuous Auditing Case Study

In the traditional audit approach, sampling selection is often done with manual intervention on Excel exports. This is a time-consuming exercise that can be further prolonged by the size of the export, its format and data structure. Manual filtering and Excel formulas for such exports may cause a crash of the program and start the process all over again. Some of it may be done with a software tool like IDEA (v. 3.12.4) for example. While using the tools is time-saving, it still requires knowledge of using the tool and its features and selecting the different types of features according to the engagement, as well as manually uploading the population datasets into the software, which can be a time-consuming exercise and still presents a risk of errors when merging is done. For more in-depth analysis and review manual work is needed as well.

The manual sample selection process contains a variety of steps.

#### 3.4.1. Period Definition

Depending on the type of engagement and the risk score of the audited process or entity, the first period for testing needs to be defined. Based on audit resources, risk and process knowledge, the period can be defined within different ranges—from days to the whole year.

#### 3.4.2. Data Extraction

After the period is defined, the auditor needs to reach out to the process owner and request that the transactions for the requested period be exported. In that step, risk is present when, for example, an auditor is not familiar with the ERP in use, and some transactions can be filtered out by the process owner or certain document types missed for review.

#### 3.4.3. Data Reconciliation and Integrity Checks

All the exports provided for further review need to be reconciled against another source, and data integrity needs to be verified. This means reconciling either to the General ledger or cross-checking against systems screenshots of the ERP in use.

#### 3.4.4. Data Cleansing

Oftentimes, not all columns are relevant for sampling and further investigation, so it is best to remove them from the export to squeeze the size of the file and to make it easier to navigate through. However, the data cleansing step described above refers to the traditional manual audit workflow only and does not form part of the ML model data preparation process, which is described in detail in Section 4.2.

#### 3.4.5. Actual Sampling

This is the most important part and the most time-consuming step. Depending on the features that the auditor wants to include for further analysis and action, the dataset must be worked on with a combination of pivot tables, conditional formatting and complex formulas to get to the specific set of transactions that will be subject to substantive testing.

Depending on the nature of business operations and value, a concentration of billing transactions to a customer, where, for example, 50% of its billing is done in the last 5 days of the month and its postings are not done by a scheduled batch, will require all various actions to be captured, and there is still a risk of missing a step in the process. The concentration percentage is to be determined by each organization's risk assessment based on volume and type of business.

To optimize the process and reduce detection risk posed by needed manual interventions, the current study aims to present how—by applying ML models based on supervised learning—organizations can achieve process optimization, time saving and risk reduction. For the purpose of the research, an actual production export was obtained containing more than 75,000 records stretching over a six-month period. Comparison of the main metrics of the models is done, which describes their potential benefits to the process.

#### 3.4.6. Data Collection

The data used to train and evaluate the ML models is extracted from a production environment by a company in the private sector. The VBRK (Billing Document: Header Data) table has been derived from SAP and contains historical billing document records. It is a standard and reliable part of enterprise SAP systems, often used in financial and compliance reporting. This structured and business-relevant data input is an essential part of building the case for training ML models with a classification task aimed at identifying unusual activity, which may present risk to the company's operating financials or gaps in existing internal control mechanisms. The VBRK table contains detailed information about each document, but only essential fields, which are indicators of possible risky or fraudulent activity, are chosen as variables to build the model.

#### 3.4.7. Key Variables

The variables are selected based on their significance in billing behavior and audit relevance, and potential predictive power. The selection is made to capture document type, destination, timing, and financial characteristics. For clarity the export contains the business name of the fields, which are different from the technical ones in a pure VBRK export. The match between both is presented in Table 2, alongside all variables used for model inputs.

**Table 2.** Key variables.

| Technical Field | Business Name       | Description                                                                                                                                  |
|-----------------|---------------------|----------------------------------------------------------------------------------------------------------------------------------------------|
| VBELN           | Billing Document    | Unique identifier of each billing transaction; enables traceability across SAP modules and linkage between financial and logistical records. |
| KALSM           | Pricing Procedure   | Configuration of pricing, discount and tax determination rules; used to identify irregular or exceptional pricings.                          |
| FKDAT           | Billing Date        | Operational date when the billing document was executed; provides the temporal anchor for anomaly detection.                                 |
| NETWR           | Net Value           | Monetary value of the billing document excluding taxes; core numeric variable for pattern and rounding analyses.                             |
| ERNAM           | Created by          | User ID that created the billing document; distinguishes manual from automated postings for behavioral insights.                             |
| ERZET           | Time of Creation    | System time of billing document creation; used to derive working-hour and off-hour activity indicators.                                      |
| ERDAT           | Creation Date       | System date on which the billing document was entered; allows comparison with billing date to detect delayed or back-dated postings.         |
| KUNRG           | Sold to Party       | Customer identifier; enables behavioral profiling and detection of customer-specific anomalies.                                              |
| STCEG           | VAT Registration N  | Customer's VAT identification number; used to verify tax compliance and identify missing or invalid registrations.                           |
| FKSTO           | Canceled            | Indicator showing whether the billing document has been canceled or reversed.                                                                |
| VSBED           | Shipping Conditions | Delivery or transportation mode analyzed for deviations from a customer's dominant shipping pattern.                                         |
| BZIRK           | Sale District       | Sales area or regional market segment; used to test structural or geographic billing consistency.                                            |

### 3.4.8. Control Variables

Some of the relevant SAP fields are retained solely as control variables for governance and contextual integrity. They are excluded from the predictive feature set due to limited discriminatory power within the available dataset; as the data originates from a single organizational entity, fields such as Company Code, Sales Organization, and Division contain only one or very few distinct values across all records, rendering them invariant and therefore non-informative for the classification task. In multi-entity or cross-organizational datasets, these fields would be expected to carry meaningful predictive signals and should be reconsidered as active model features accordingly. Those are treated as invariant rather than as a predictive driver in the model. The list of invariants is presented in Table 3. For different organizations or available datasets, those can be key components in the model input data, depending on the dataset format and identified risks for the organization.

**Table 3.** Control variables.

| Technical Field | Business Name      | Description                                                                                                                                                                                       |
|-----------------|--------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| VKORG           | Sales Organization | Organizational unit responsible for sales operations; reflects structural or regional differences influencing billing behavior.                                                                   |
| FKART           | Billing Type       | Defines the type of billing document (e.g., invoice, credit memo, debit memo, pro forma); determines document behavior, number range, and posting logic within the financial and sales processes. |

Table 3. *Cont.*

| Technical Field | Business Name         | Description                                                                                                                                                                                               |
|-----------------|-----------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| FKTYP           | Billing Category      | Higher-level classification grouping different billing; useful for differentiating document purposes in audit and reporting contexts.                                                                     |
| VTWEG           | Distribution Channel  | Specifies the channel through which goods or services reach the customer (e.g., direct sales, wholesale, online); important for analyzing channel-specific billing irregularities and pricing structures. |
| BUKRS           | Company Code          | Legal accounting entity within SAP; defines the set of books and fiscal control area for financial postings; essential for cross-company comparisons and internal-control segmentation.                   |
| SPART           | Division              | Represents a product or service line within the organization (e.g., electronics, maintenance, software); supports stratified analysis of billing behavior across business units.                          |
| LAND1           | Destination Country   | Country to which goods or services are billed or shipped; relevant for tax compliance and export validation.                                                                                              |
| MWSKZ/TAXD1     | Tax Departure Country | Country of origin for tax determination; relevant for cross-border VAT and intercompany billing checks.                                                                                                   |

### 3.4.9. Engineered Audit-Related Variables

The engineered audit-relevant variables extend the raw SAP billing data with domain-specific indicators that are used for capturing behavioral and temporal irregularities.

These derived features capture audit-critical dimensions—such as unusual pricing patterns, weekend or off-hour billing, missing VAT information, or deviations from dominant process behavior—and thus enhance the interpretability and governance value of the predictive model.

All audit-engineered features are defined as numeric features in the model definition. Those features are set to be key parameters for the model data definition and as such play a key role in determining each risky class.

The audit-engineered features are described in detail in Table 4.

Table 4. Audit-engineered variables.

| Name                          | Description                                                                                                                                                                     |
|-------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Weekend Day (WD)              | Binary indicator for whether the billing or creation date occurred on a weekend, potentially signaling irregular posting practices.                                             |
| Out-of-Range Hour (ORH)       | Binary flag identifying documents created outside standard working hours (before 07:00 or after 19:00).                                                                         |
| Manual Flag (MF)              | Indicates that the document was created manually (not by auto or batch users).                                                                                                  |
| Off-Hours Manual (OHM)        | Combined flag marking manually created documents outside working hours—often correlated with riskier billing behavior.                                                          |
| Net Value Modulo (netwr_mod1) | Represents the decimal remainder of the transaction's net value; this variable is designed to capture potential rounding patterns (amounts ending in 0.99, 0.95, 0.49 or 0.00). |
| Suspicious Decimal (SD)       | Binary flag for patterned amounts ending with common marketing or psychological pricing digits (e.g., 0.99).                                                                    |
| Edge-of-Month Window (EOMW)   | Indicates whether a billing date falls within the first or last five days of a month.                                                                                           |

Table 4. Cont.

| Name                       | Description                                                                                                                    |
|----------------------------|--------------------------------------------------------------------------------------------------------------------------------|
| Party Edge-Prone (PEP)     | Identifies customers whose billing behavior systematically clusters near month-end cut-offs (>50% of invoices in EOM windows). |
| Over 85 K Including (O85K) | Marks customers whose cumulative yearly net purchases exceed €85,000.                                                          |
| VAT Missing Flag (VAT_MF)  | Binary variable capturing missing or invalid VAT registration numbers per Sold-To party.                                       |

### 3.5. Classification Categories

For the purposes of the presented study, we introduced a total of five classes that can be considered generally risky in audit and are of interest for both internal and external auditors:

- Class 1 (0.99; 0.95; 0.49; 0.00)—Patterned amount: In the audit world, documents ending with those numbers after the decimal may catch the attention of auditors because they represent a specific risk. Those patterned numbers, for example, 0.99, may fall below a determined threshold for automated postings, billings or payments. Round numbers, on the other hand, also represent a certain level of risk as they are a feature of fraudulent activities and manipulation of document creation, which can be hard to draw attention to in large sets of data. They stand out when analyzing invoice data for anomalies since genuine transactions typically show varied and precise amounts due to taxes, discounts, and other adjustments. Additionally, frequent or unexplained round-number transactions may be a signal of bypassing of internal controls to hide unauthorized payment.
- Class 2 Creations outside usual working hours activity, created manually (weekend activity, or outside normal working hours (defined as before 07:00 or after 19:00 local system time)). However, this definition can be adjusted by each organization depending on its operational structure and business hours. Those transactions are flagged under the assumption that activity outside of normal working hours and days requires extra security for possible policy violations or other fraudulent activity. The majority of control mechanisms, review procedures and supervisory activities are designed for standard business hours. Transactions at unusual times are classic red flags in anomaly detection models, flagging behavior outside an entity's typical operational patterns.
- Class 3 Anomaly in shipping conditions of a customer (different shipping condition different from the main one (main >65% and more than 3 invoices))—the reason for flagging is that it is a signal of fraud and because it may indicate attempts to bypass standard controls, manipulate delivery terms or disguise the true nature of transactions. Shipping conditions, such as expedited shipping, special handling, or altered delivery terms, affect pricing, costs and timing. Unusual or inconsistent use that deviates from normal customer patterns can raise suspicion.
- Class 4 Concentration of documents close to month end (>50% in the last and first five days of total documents crated for a specific month)—A concentration of invoices close to month-end or at the beginning of a new month represents an audit risk primarily due to the potential for financial statement manipulation and errors during the busy and pressured month-end close process and trying possibly to boost revenue or undermine expenses. Invoices dated close to period-end may not align with the actual delivery or service dates, leading to misstatements in financial period closing activities.
- Class 5 Missing VAT on customers after reaching a cumulative amount >85,000 EUR for the financial period—such documents are flagged under the assumption that missing VAT can be a signal of concerns about tax compliance, potential VAT fraud and the legitimacy of the business relationship. As the production dataset is in a

jurisdiction where the regulation for VAT registration is once 85,000 turnovers from goods is reached, the flag setup is required. On the other hand, missing VAT generally can indicate attempts to evade VAT rules, claim incorrect VAT refunds, or participate in missing trader or carousel fraud cases where taxes are collected but not paid to authorities.

- **Class 0—Regular Billing:** documents that do not trigger any of the five defined audit-relevant anomaly criteria described above. These transactions do not carry detectable indicators of the flagged risk patterns and are therefore classified as compliant within the scope of this study. Class 0 serves as the baseline majority class, representing standard billing behavior against which anomalous classes are distinguished. Those are considered non-risky in the context of the current study.

By flagging those based on a full population and with the possibility of near real-time flags, internal auditors will have a convincing case for follow-up activities to ensure all errors are corrected if present. In this way, the possibility of fraud will be limited, and internal controls will be operating as intended.

For better results interpretation and avoiding duplication, which may cause issues with ML model output and delay any timely follow-up investigation, a different weight is assigned to each class. When a transaction is flagged in one class, it will not appear in the rest if it shares more than one risky feature. In the current study, the highest priority is assigned in class 2—activity outside of regular business hours, meaning that if a document is created on Sunday, ending at 0.99 or with a missing VAT, it will be flagged in class 2 and will not appear in the rest of the defined classes.

The assigned weight is as follows:  $2 > 4 > 3 > 5 > 1 > 0$ . Of course, that weight can be rearranged to an individual organization's requirements based on their ERP and risk assessment procedures.

## 4. Multinomial Logistic Regression Model for Continuous Auditing

### 4.1. Model Selection

Multinomial logistic regression is selected as the core classification algorithm in this study. The model selection is further supported by its proven performance on a publicly available Kaggle dataset used in another study (Duhova et al. 2025). The choice reflects the model's interpretability, robustness and suitability for auditing and financial risk contexts. Unlike black-box approaches such as neural networks or ensemble models, logistic regression allows for a transparent examination of how individual features contribute to class decisions—an essential requirement in internal audit and governance environments. Those features of the model are relevant to the possible need for a follow-up audit of the model itself.

Driven by the EU AI Act, which is effective from August 2025, the act requires transparency on model architecture and training data, rigorous risk assessment, conformity checks, copyright documentation, and internal audit trails.

The model is designed to classify billing documents across multiple audit-relevant anomaly categories rather than relying on a single binary fraud flag. The five defined anomaly classes represent distinct audit dimensions, including patterned monetary amounts, working-time irregularities, shipping-condition deviations, edge-of-month behaviors, and missing VAT information after yearly thresholds. The multinomial formulation of logistic regression enables probabilistic differentiation among these categories, offering a structured and explainable foundation for risk-based decision-making in SAP billing processes.

#### 4.2. Data Preparation and Filter Engineering

The dataset used for this study originates from an SAP VBRK-based billing extract representing finalized transactions from a live production environment that originated with more than 75,000 records. To ensure analytical validity, all canceled or incomplete records were excluded, leaving only 69,158 financially relevant billing documents that accurately reflect operational business activity.

Subsequent data preparation focused on harmonizing and transforming key SAP attributes related to document metadata, organizational dimensions, pricing structures, and timestamps. Empty string values were standardized as missing, numeric and datetime fields were correctly typed and date columns were enriched with weekday and edge-of-month indicators to capture temporal irregularities. Text-based attributes such as Created by and VAT Registration No. were normalized to a uniform format to enable pattern-based analysis.

From this cleansed foundation, a set of engineered audit-relevant variables was derived to capture behavioral and compliance-oriented signals linked to SAP billing logic. These include indicators for suspicious amount patterns (e.g., prices ending in 0.99 or 0.00), manual activities outside standard working hours, deviations in customer-specific shipping conditions, edge-of-month billing concentration, and missing VAT information once the statutory annual threshold ( $\geq \text{€}85,000$ ) is exceeded.

Each of these engineered features corresponds to an audit scenario and forms the basis of five anomaly classes—ranging from pricing and timing irregularities to compliance risks. A hierarchical rule ensures mutual exclusivity between classes, while standard scaling and one-hot encoding prepare the data for multinomial logistic regression. The resulting dataset thus combines transactional transparency with analytical depth, aligning SAP operational semantics with machine-learning-based anomaly detection for audit and governance purposes.

#### 4.3. Feature Analysis

In audit-oriented machine learning, model transparency and explainability are important prerequisites for acceptance and operational deployment. To ensure interpretability, this study incorporates a structured feature influence analysis that quantifies how each variable contributes to the prediction of a specific anomaly class. This approach allows the transformation of statistical weights into meaningful audit insights, for example, identifying which behavioral indicators most strongly drive the detection of off-hour billing, missing VAT information or atypical shipping behavior.

Feature influence was visualized through feature-to-class influence distribution, where each feature's total contribution is distributed across anomaly classes in percentage terms, as shown in Figure 2. The visualization shows how much each feature contributes to determining classes. Higher values indicate a stronger predictive influence of the respective features for the associated class. The figure suggests correct choices for the key variables as, for example, weekend day and offhours\_manual are the key influential factors to Class 2—Creations outside of usual working hours activities. Another example is VAT\_missing\_flag and over\_85k\_including as they are the main influencers on class 5. The rest of the features are in line with the selection of classes and contribute on a very good basis to the flagged documents in selected classes. By expressing coefficients as normalized proportions, the model's interpretive structure becomes intuitive even to non-technical audit professionals, supporting cross-disciplinary review and governance.

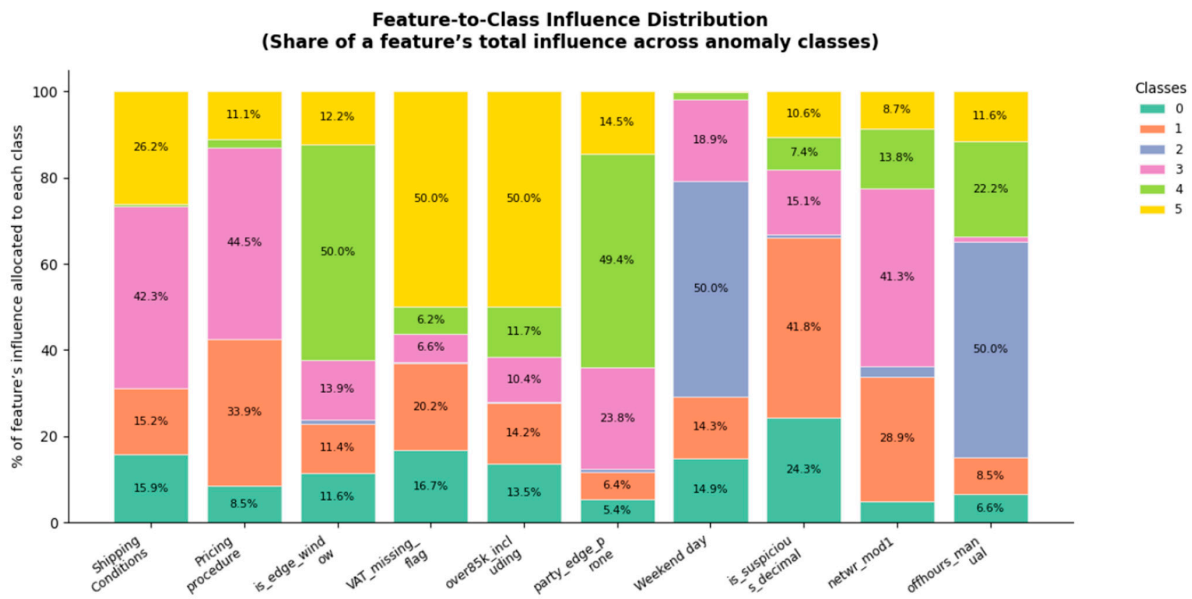


Figure 2. Feature to class influence distribution.

In addition, a feature correlation heatmap is extracted to assess interdependencies between the numerical features, ensuring that redundant or collinear features do not distort interpretive weight, as shown in Figure 3. The correlation matrix visualizes the linear relationships among the engineered numeric audit features used as predictors in the logistic regression model. The coefficients range from  $-1$  (perfect negative correlation) to  $+1$  (perfect positive correlation).

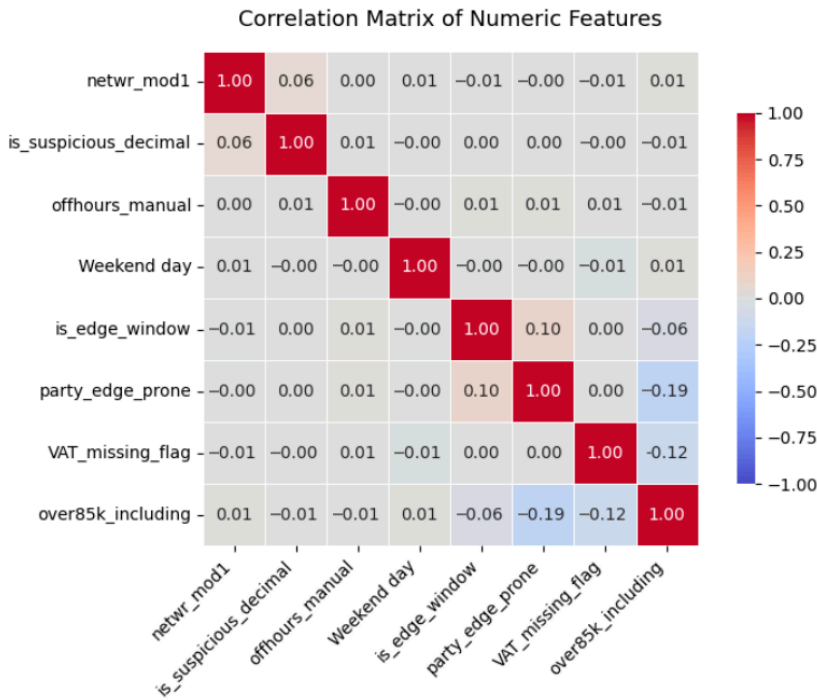


Figure 3. Feature correlation heatmap.

As the heatmaps show, there is very low correlation—nearly all coefficients fall between  $-0.20$  and  $+0.10$ , indicating minimal collinearity between engineered features. This is an excellent result—it means that the engineered indicators (e.g., `offhours_manual`, `is_edge_window`, `VAT_missing_flag`) capture distinct audit behaviors rather than overlap-

ping patterns. In addition, each variable contributes unique information to the classifier and enhances model interpretability.

There are mild isolated correlations visible on the heatmap. A small positive correlation ( $\sim 0.10$ ) appears between `is_edge_window` and `party_edge_prone`, which is expected: customers frequently invoicing near month-end tend to create clusters of invoices during closing days. Another example is the weak negative correlation ( $-0.19$ ) between `party_edge_prone` and `over85k_including`, which may indicate that edge-of-month billing is slightly less common among high-value or VAT-threshold customers. This possibly reflects different operational cycles or tax-driven constraints, which can be subject to additional analysis depending on the desired outcome. Core behavioral indicators such as `offhours_manual`, `is_suspicious_decimal` and `VAT_missing_flag` show almost zero correlation with temporal or value-based features. Such independence is desirable for both multinomial classification and audit explainability, as each anomaly class corresponds to a unique behavioral signature.

Together, these explainability methods provide an audit-aligned perspective on machine learning, where each model outcome can be traced back to operational logic, business process behavior or policy compliance thresholds. This enables not only anomaly detection but also knowledge generation about the underlying process dynamics within SAP billing workflows.

The correlation matrix of the engineered audit variables demonstrates very low inter-feature dependence, with all coefficients remaining within the  $-0.2 \leq r \leq 0.1$  range. This confirms that the derived features capture distinct behavioral dimensions of billing activities such as temporal irregularities, value rounding patterns and VAT compliance deviations—without introducing multicollinearity.

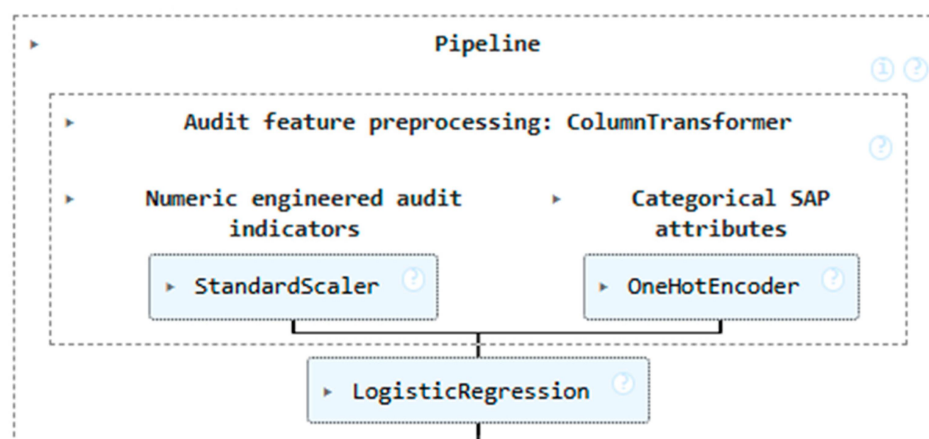
The mild correlation between `is_edge_window` and `party_edge_prone` reflects expected temporal co-occurrence, while all other relationships remain statistically negligible. This structural independence supports both the interpretability and numerical robustness of the multinomial logistic regression model, ensuring that each audit-relevant feature contributes uniquely to anomaly detection and risk classification.

#### 4.4. Model Training and Evaluation

The multinomial logistic regression model was trained using a stratified 80/20 train-test split to ensure proportional representation of each anomaly class across both subsets. A unified preprocessing and modeling pipeline was implemented to guarantee consistency across the full analytical workflow. The pipeline combines feature transformation and model fitting into a single reproducible structure.

**Implementation and reproducibility.** To ensure consistent preprocessing and fully reproducible inference, the model was implemented as a single scikit-learn Pipeline that integrates feature transformation and classification. Stratified sampling was applied by using `stratify = y` during the train-test split to preserve the relative class proportions across all subsets, ensuring representative evaluation of rare anomaly categories.

Numerical engineered audit indicators were standardized using `StandardScaler`, while categorical SAP attributes were encoded using `OneHotEncoder` with unknown-category handling, followed by multinomial logistic regression (Softmax). This pipeline structure guarantees that identical transformations are applied during both training and evaluation, reduces the risk of manual preprocessing errors, and supports auditability through an explicit, inspectable model workflow. Figure 4 illustrates the reproducible preprocessing and classification pipeline for audit anomaly detection.



**Figure 4.** Pipeline for audit anomaly detection.

#### 4.5. Dataset Distribution

The dataset comprises 69,158 finalized billing records, partitioned into 55,326 training (80% of all data) and 13,832 testing (20% of all data) instances, preserving the same proportional representation across both subsets, as shown in Table 5.

The Regular Billing (Compliant Transactions) class dominates the dataset with  $\approx 89\%$  of all records, reflecting the naturally imbalanced structure of real-world enterprise billing data, where the majority of transactions are compliant.

The remaining  $\approx 11\%$  are distributed among five audit-relevant anomaly classes:

- Patterned Amount (4.3%) and Missing VAT (4.7%) represent the largest anomalous segments, suggesting frequent value-pattern and tax-compliance irregularities;
- Edge-of-Month Behavior (1.1%) occurs moderately often, typically linked to period-end adjustments;
- Shipping Condition Anomalies (0.4%) and Working-Time Anomalies (0.2%) are rare but high-risk categories, indicating potential process deviations rather than volume-based irregularities.

**Table 5.** Dataset distribution.

| Class ID | Class Name           | Train N | Train % | Test N | Test %  | Total N |
|----------|----------------------|---------|---------|--------|---------|---------|
| 0        | Regular billing      | 49,387  | 89.27%  | 12,347 | 89.27%  | 61,734  |
| 1        | Patterned amount     | 2364    | 4.27%   | 591    | 4.27%   | 2955    |
| 2        | Working time anomaly | 134     | 0.24%   | 33     | 0.24%   | 167     |
| 3        | Shipping anomaly     | 200     | 0.36%   | 50     | 0.36%   | 250     |
| 4        | Edge of the month    | 630     | 1.14%   | 158    | 1.14%   | 788     |
| 5        | Missing VAT          | 2611    | 4.72%   | 653    | 4.72%   | 3264    |
| Total    |                      | 55,326  | 100.00% | 13,832 | 100.00% | 69,158  |

#### 4.6. Evaluation and Reporting of the Logistic Regression

The logistic regression classifier was configured in multinomial mode with the lbfgs solver and class weights balanced inversely to class frequency. This approach addresses class imbalance and prevents overfitting to the dominant Regular Billing (Class 0) category. The Softmax activation ensures that the model outputs a complete probability distribution across all classes, allowing risk scoring and audit prioritization to be expressed probabilistically rather than deterministically.

To ensure classification fairness, both argmax-based classification and probability thresholding ( $\theta$ ) were evaluated. As demonstrated below, both approaches yielded equivalent performance metrics, confirming that argmax classification provides a fair and reliable

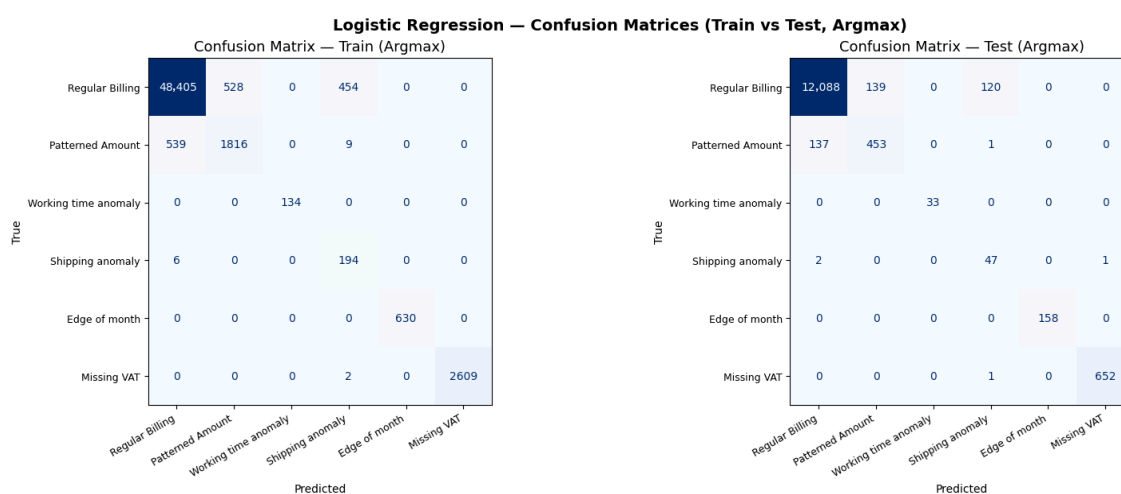
basis for result reporting in this study. The argmax approach assigns each record to the class with the highest predicted probability, providing a direct view of the model's inherent decision boundaries. The threshold-based approach introduces a tunable confidence level ( $\theta = 0.8$ ), where only predictions exceeding the defined probability threshold are classified into an anomaly class. Otherwise, the observation defaults to Regular Billing. This hybrid evaluation provides transparency into model certainty, as shown in Table 6, allowing auditors to calibrate sensitivity according to organizational risk tolerance.

**Table 6.** Threshold performance comparison of the logistic regression classifier.

| Split/Decision        | Accuracy | Precision (Macro) | Recall (Macro) |
|-----------------------|----------|-------------------|----------------|
| Train—Argmax          | 0.972    | 0.843             | 0.953          |
| Test—Argmax           | 0.971    | 0.838             | 0.947          |
| Train— $\theta = 0.8$ | 0.973    | 0.847             | 0.951          |
| Test— $\theta = 0.8$  | 0.972    | 0.843             | 0.947          |

Following the re-engineering of features and audit-specific variable integration, the model's class separation improved notably, resulting in equivalent performance between the standard argmax classification and the probability-threshold approach ( $\theta = 0.80$ ). Therefore, only argmax-based results are presented for clarity.

Figure 5 presents the confusion matrices for both the training and test datasets using the argmax classification strategy.



**Figure 5.** Confusion matrices for training and test data for the logistic regression classifier.

The model demonstrates strong class separation and stable generalization performance across all six target categories. The dominant diagonal elements indicate a high proportion of correctly classified samples in each class. The majority class, Regular Billing (Class 0), achieves nearly perfect accuracy, with over 98% of its instances correctly identified in both datasets. Similarly, the minority anomaly classes—Working-time anomaly (Class 2), edge-of-month (Class 4) and Missing VAT (Class 5)—are also classified with perfect recall (100%) and no confusion with other categories, confirming that the feature-engineering process effectively captured their distinctive behavioral patterns.

The Patterned Amount (Class 1) class maintains solid performance ( $\approx 77\%$  precision and recall), showing that the model reliably detects transactions with repetitive or structured amount patterns while allowing minor overlap with normal cases. The Shipping anomaly (Class 3) class exhibits comparatively lower precision ( $\approx 0.28$ ) but high recall ( $\approx 0.94$ ), meaning that while most true shipping anomalies are detected, some regular

records are incorrectly assigned to this class. This asymmetry suggests that the feature representation of this class still partially overlaps with regular or patterned billing behavior—a typical issue in imbalanced anomaly detection tasks.

Importantly, the consistency between the Train and Test confusion matrices confirms that the model generalizes well without overfitting. The near-identical patterns across both datasets reflect the success of the stratified data split and the inclusion of engineered audit features, which improved discriminatory power across all anomaly categories.

Performance metrics were computed across both methods, including Accuracy, Precision, Recall and F1-score. Classification report for train and test sets was analyzed to assess per-class separability and potential overfitting. A visualization of the results follows in Table 7.

**Table 7.** Classification report for the logistic regression classifier.

| Class ID            | Precision Train | Precision Test | Recall Train | Recall Test | F1-Score Train | F1-Score Test | Train Set N | Test Set N |
|---------------------|-----------------|----------------|--------------|-------------|----------------|---------------|-------------|------------|
| 0                   | 0.99            | 0.99           | 0.98         | 0.98        | 0.98           | 0.98          | 49,387      | 12,347     |
| 1                   | 0.78            | 0.77           | 0.77         | 0.77        | 0.78           | 0.77          | 2364        | 591        |
| 2                   | 1.00            | 1.00           | 1.00         | 1.00        | 1.00           | 1.00          | 134         | 33         |
| 3                   | 0.29            | 0.28           | 0.97         | 0.94        | 0.45           | 0.43          | 200         | 50         |
| 4                   | 1.00            | 1.00           | 1.00         | 1.00        | 1.00           | 1.00          | 630         | 158        |
| 5                   | 1.00            | 1.00           | 1.00         | 1.00        | 1.00           | 1.00          | 2611        | 653        |
| <b>Accuracy</b>     | 0.97            | 0.97           | 0.97         | 0.97        | 0.97           | 0.97          | -           | -          |
| <b>Macro Avg</b>    | 0.84            | 0.84           | 0.95         | 0.95        | 0.87           | 0.86          | 55,326      | 13,832     |
| <b>Weighted Avg</b> | 0.98            | 0.98           | 0.97         | 0.97        | 0.97           | 0.97          | 55,326      | 13,832     |

The results in Table 7 show a consistently high level of accuracy (97%) and strong alignment between train and test scores across all metrics, indicating stable generalization and minimal overfitting.

The majority class (Regular Billing) maintains excellent precision and recall (0.99/0.98), confirming that normal transactional patterns are accurately distinguished from anomalies. Among the anomaly classes, Patterned Amount achieves balanced precision and recall of 0.77, demonstrating the model’s capability to detect repetitive or structured amount patterns with moderate overlap to regular billing.

Working-time anomaly—associated with transactions posted outside of the defined working hours—edge-of-month, and Missing VAT are all classes that achieve perfect F1-scores (1.00), confirming that the engineered temporal and tax-related features effectively isolate their distinctive patterns.

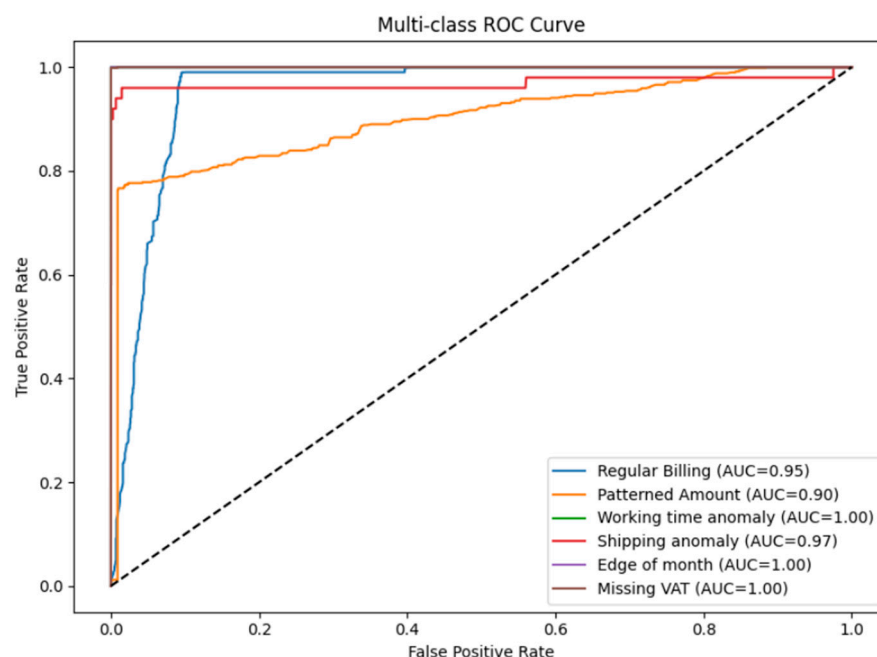
The Shipping anomaly class remains the most challenging to separate, with a precision of 0.28 and a recall of 0.94, suggesting that while most true anomalies are captured, a notable fraction of regular or patterned cases is misclassified as shipping anomalies. This reflects the feature overlap identified in the confusion matrix and indicates a potential area for targeted feature refinement (e.g., inclusion of delivery-term or shipping-region metadata).

The macro-averaged F1-score (0.86) demonstrates balanced performance across classes despite dataset imbalance, while the weighted F1-score (0.97) confirms the model’s robustness in production-scale conditions dominated by legitimate billing records. The close correspondence between training and test metrics across all measures reinforces the effectiveness of the stratified sampling strategy and the inclusion of engineered audit features, which jointly improved class discrimination and model interpretability.

#### 4.7. Model Performance on Test Data

Receiver Operating Characteristic (ROC) and Precision–Recall (PR) curves were generated for each anomaly class, providing a visual understanding of the trade-off between sensitivity and specificity only for the test data.

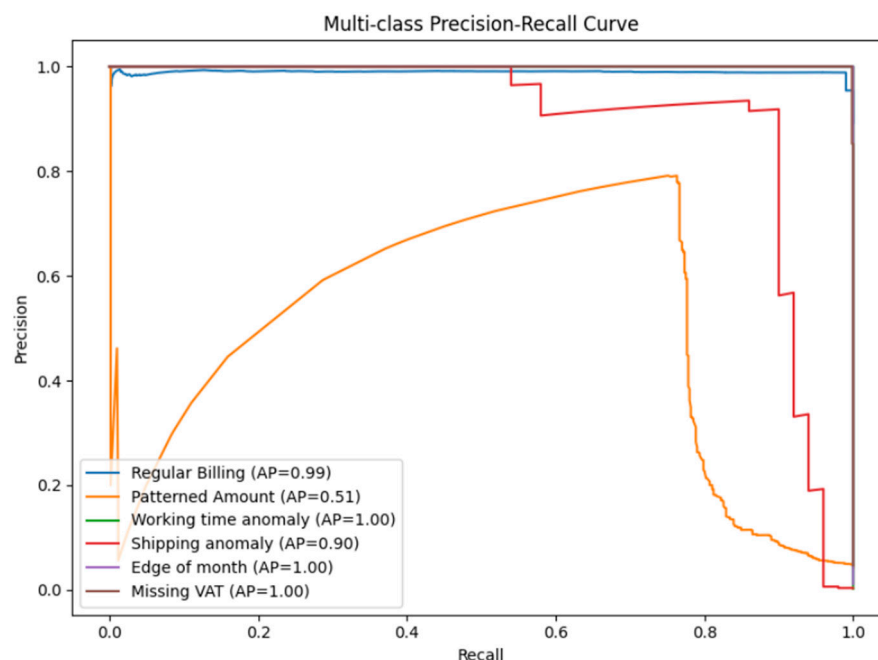
The multi-class ROC analysis, present in Figure 6, further confirms the robustness and discriminative capability of the trained classifier across all categories. The Area Under the Curve (AUC) values remain consistently high, indicating that the model achieves strong separability between normal and anomalous billing behaviors.



**Figure 6.** Multi-class ROC curve.

- Regular Billing (AUC = 0.95) shows excellent true-positive discrimination with minimal false alarms, validating that the model reliably identifies legitimate transactions.
- Patterned Amount (AUC = 0.90) demonstrates slightly lower performance compared to the other classes. This reflects the subtle overlap between recurring or threshold-based amounts and legitimate billing, where numeric rounding and business rules can mimic anomaly signals.
- Working-time anomaly (AUC = 1.00), Edge of month (AUC = 1.00) and Missing VAT (AUC = 1.00) each achieved perfect separability, confirming the effectiveness of the engineered temporal and tax-related features in isolating these irregularities.
- Shipping anomaly (AUC = 0.97) also performs strongly, though its slightly lower precision (as seen in the classification report) indicates some residual confusion with other operational edge cases, which is typical in multi-class business datasets with overlapping process conditions.

As can be seen in this section, the ROC curves collectively illustrate high sensitivity and specificity across all classes, with macro-AUC  $\approx 0.97$ , suggesting that the model maintains consistent decision boundaries without bias toward the dominant Regular Billing class. These results corroborate the precision–recall findings, shown in Figure 7, and confirm that the feature-engineered model generalizes effectively across diverse anomaly patterns in the audit dataset.



**Figure 7.** Precision–Recall curve.

The multi-class Precision–Recall analysis complements the ROC evaluation by emphasizing the trade-off between precision (false-positive control) and recall (sensitivity) in a highly imbalanced classification context. The model demonstrates strong class separability and robust predictive confidence for most categories, which are supported by high average precision (AP) scores.

- Regular Billing (AP = 0.99) exhibits near-perfect precision–recall behavior, confirming that the model reliably identifies normal cases with minimal misclassification.
- Patterned Amount (AP = 0.51) performs less consistently. The moderate AP value indicates that this class remains challenging, likely due to overlapping numerical patterns with legitimate billing records—an inherent difficulty in detecting subtle rounding or pricing regularities in enterprise data.
- Working-time anomaly (AP = 1.00), Edge of month (AP = 1.00) and Missing VAT (AP = 1.00) show perfect performance, verifying that temporal and VAT-related engineered features provide strong signal clarity and allow for complete precision–recall alignment.
- Shipping anomaly (AP = 0.90) maintains a high AP but reflects a few borderline cases where anomalies resemble operational irregularities from other categories, consistent with their slightly lower precision observed earlier.

Taken together, the PR curves indicate high discriminative confidence across minority classes, with average precision values closely mirroring the AUC scores. The clear separation of curves and sharp ascents to high precision at early recall levels demonstrates that the classifier maintains strong reliability even in the presence of class imbalance—a crucial factor for risk-based auditing and continuous control monitoring scenarios.

The logistic regression model achieved stable generalization across training and test datasets, with interpretable class contributions aligned with known audit risks. However, for some of the risk classes, it struggles to classify them with high levels of precision.

For that reason, another type of ML model was applied to the same dataset in order to compare its results to logistic regression and create an overview of both types of models and their respective outputs.

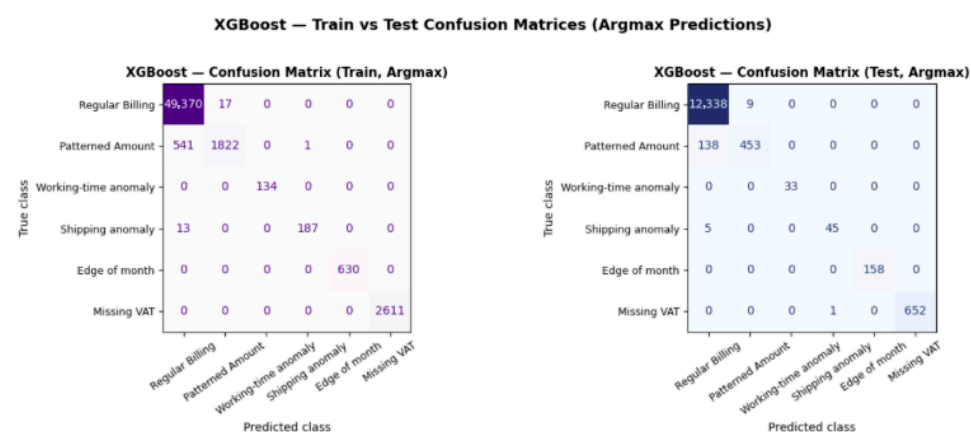
## 5. XGBoost Model for Continuous Auditing

### 5.1. Dataset Distribution

To ensure a consistent basis for comparison with the multinomial logistic regression baseline, the XGBoost classifier was trained and evaluated using the same train–test split. The training subset contains 55,326 billing records, while the test subset contains 13,832 records. The resulting label distribution is highly imbalanced, reflecting real-world ERP billing populations, in which compliant transactions dominate, and anomalous or audit-relevant events occur at substantially lower frequency. This imbalance is methodologically important because overall accuracy can be inflated by the majority class. Therefore, the evaluation emphasizes class-specific precision and recall, confusion-matrix diagnostics, and probability-based ROC and precision–recall (PR) analyses, which are more informative for audit deployment scenarios.

### 5.2. Evaluation and Reporting of the XGBoost Model

Model performance was assessed using argmax-based multi-class predictions and reported through confusion matrices for the training and test datasets and per-class classification report, including precision, recall, and F1-score. Results are shown in Figure 8. This reporting structure is aligned with audit-oriented evaluation needs because it exposes not only aggregate performance but also the operational error pathways. In continuous auditing, such error pathways determine whether the model primarily generates false alerts (low precision) or misses relevant exceptions (low recall), both of which directly affect investigation workload and residual risk.



**Figure 8.** Confusion matrices for training and test data for XGBoost classifier.

The confusion matrices demonstrate strong class separation and stable generalization, visible in Figure 8. Regular Billing is classified almost perfectly in both datasets, with 49,370/49,387 correct predictions in training and 12,338/12,347 in test, indicating that the model does not incorrectly shift large volumes of normal transactions into anomaly classes. This property is critical for operational acceptability, since excessive false positives can undermine trust and reduce adoption by audit teams.

For the anomaly classes, most misclassifications follow intuitive patterns. The Patterned Amount class exhibits the most consistent overlap with Regular Billing, with 541 training instances and 138 test instances predicted as regular. This behavior occurs because some pricing patterns can occur legitimately due to rounding rules, discount structures, or standardized pricing conventions, which can reduce separability unless additional contextual variables are available. Shipping anomaly shows minor leakage primarily into the regular class (13 instances in training and 5 instances in test), suggesting that the class boundary is generally coherent but may benefit from additional logistics-related descriptors

if the objective is further recall improvement. In contrast, Working-time anomaly and Edge of month are classified without visible confusion in the shown matrices, indicating that temporal indicators provide highly distinctive signals. Missing VAT is also classified with near-perfect reliability, with only negligible confusion on the test set.

The classification report, presented in Table 8, corroborates these findings. Overall accuracy is 0.99 on both training and test datasets, indicating high predictive quality and limited overfitting. Test-set precision is close to 1.00 across classes, including 0.99 for Regular Billing and approximately 0.98 for Patterned Amount and Shipping anomaly. Recall is also high across classes, with the main reduction observed for Patterned Amount (test recall approx. 0.77, F1 approx. 0.86). From an audit perspective, this profile is operationally favorable: when the model flags a transaction as Patterned Amount, it is usually correct (high precision), while some borderline cases remain difficult to detect without additional corroborating evidence. Importantly, macro-averaged performance remains strong (macro recall approx. 0.94 on test), which confirms that minority classes are handled effectively and that results are not dominated solely by the majority class.

**Table 8.** Classification report for XGBoost classifier.

| Class ID            | Precision Train | Precision Test | Recall Train | Recall Test | F1-Score Train | F1-Score Test | Train Set N | Test Set N |
|---------------------|-----------------|----------------|--------------|-------------|----------------|---------------|-------------|------------|
| 0                   | 0.99            | 0.99           | 1.00         | 1.00        | 0.99           | 0.99          | 49,387      | 12,347     |
| 1                   | 0.99            | 0.98           | 0.77         | 0.77        | 0.87           | 0.86          | 2364        | 591        |
| 2                   | 1.00            | 1.00           | 1.00         | 1.00        | 1.00           | 1.00          | 134         | 33         |
| 3                   | 0.99            | 0.98           | 0.94         | 0.90        | 0.96           | 0.94          | 200         | 50         |
| 4                   | 1.00            | 1.00           | 1.00         | 1.00        | 1.00           | 1.00          | 630         | 158        |
| 5                   | 1.00            | 1.00           | 1.00         | 1.00        | 1.00           | 1.00          | 2611        | 653        |
| <b>Accuracy</b>     | 0.99            | 0.99           | 0.99         | 0.99        | 0.99           | 0.99          | -           | -          |
| <b>Macro Avg</b>    | 1.00            | 0.99           | 0.95         | 0.94        | 0.97           | 0.97          | 55,326      | 13,832     |
| <b>Weighted Avg</b> | 0.99            | 0.99           | 0.99         | 0.99        | 0.99           | 0.99          | 55,326      | 13,832     |

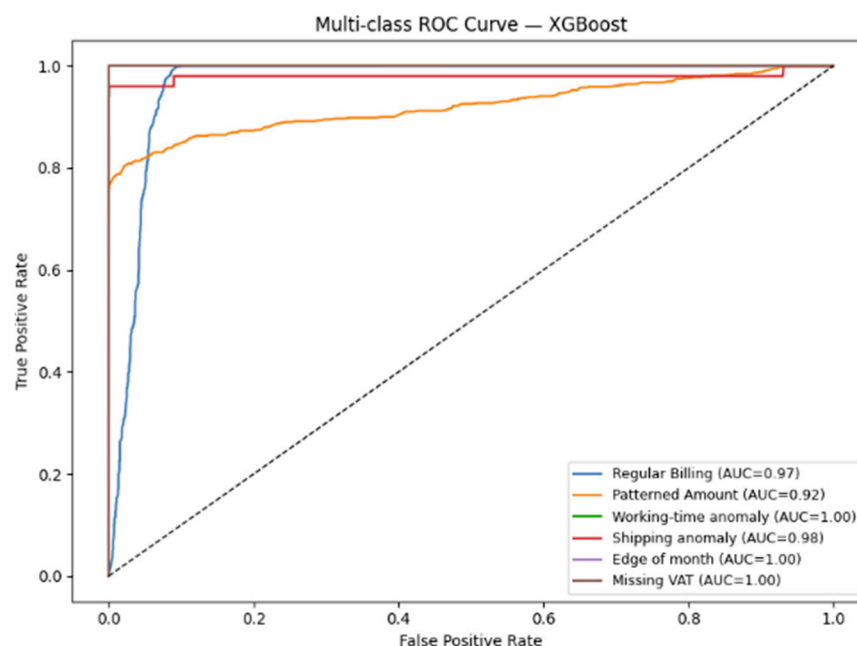
### 5.3. XGBoost Model Performance on Test Data (ROC and Precision–Recall)

To complement argmax-based classification, the model was further evaluated using probability-based performance curves on the test dataset. Multi-class ROC curves were used to assess overall ranking and separability (Figure 9), while precision–recall curves were used to evaluate performance under class imbalance (Figure 10), where PR analysis is typically more indicative of practical alert quality.

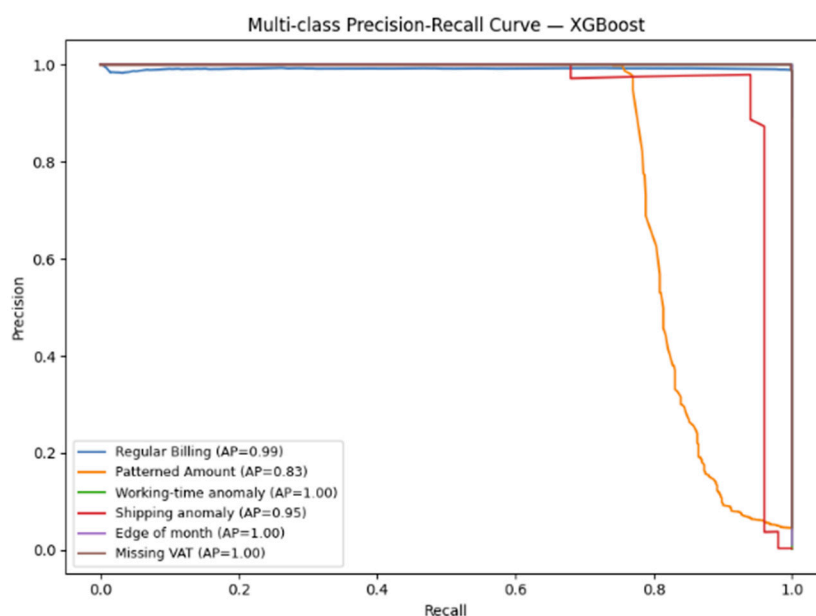
The ROC analysis, present in Figure 9, indicates excellent discriminative capability across classes. The AUC values are consistently high, with Regular Billing (AUC  $\approx$  0.97) and Patterned Amount (AUC  $\approx$  0.92) representing the most challenging separation, while Shipping anomaly (AUC  $\approx$  0.98) remains very strong. Notably, Working-time anomaly, Edge of month, and Missing VAT achieve AUC  $\approx$  1.00, suggesting near-perfect separability for these categories. This result aligns with audit interpretability expectations, since temporal and compliance indicators tend to be structurally distinctive in transactional ERP logs.

The PR analysis in Figure 10 provides an operational view of alert reliability. Average Precision (AP) is very high for most classes, including Regular Billing (AP  $\approx$  0.99), Shipping anomaly (AP  $\approx$  0.95), and near-perfect values (AP  $\approx$  1.00) for Working-time anomaly, Edge of month, and Missing VAT. The lowest PR performance is again observed for Patterned Amount (AP  $\approx$  0.83), consistent with the conceptual overlap between potentially suspicious numeric endings and legitimate pricing practices. In practical audit deployment terms, these findings suggest that XGBoost can generate reliable alerts for temporality-driven and compliance-driven anomaly types, while Patterned Amount alerts may require

complementary contextual checks (for example, pricing procedure context, customer policy, or threshold logic) to reduce residual ambiguity.



**Figure 9.** ROC for XGBoost classifier.



**Figure 10.** PR for XGBoost classifier.

The combined confusion matrix, classification report, ROC, and PR evidence demonstrates that the XGBoost model generalizes well to unseen data and provides high-quality discrimination for the anomaly taxonomy used in this study. The results support its suitability for continuous auditing workflows, where accurate prioritization and low-noise alerting are critical to maintain sustainable investigation capacity and effective risk coverage.

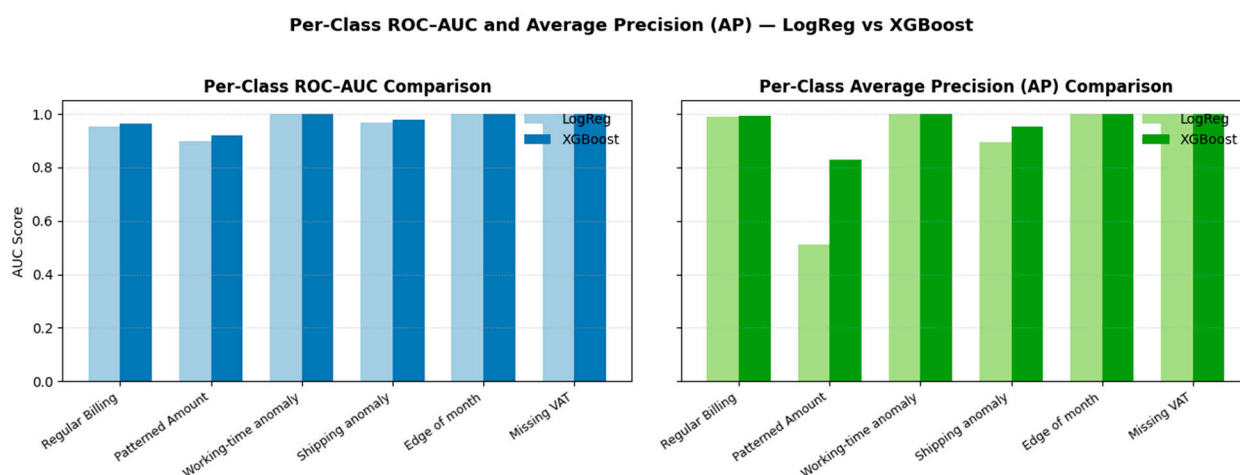
## 6. Discussion

The objective of this study was to evaluate whether supervised machine learning models can support continuous auditing by reliably flagging high-risk billing transactions extracted from ERP systems while remaining interpretable, and follow-up procedures are

done by audit departments and by integrating such an approach to assist with overall enterprise risk management into organizations. Two supervised multi-class classifiers were examined—multinomial logistic regression (SoftMax regression) and XGBoost. Their comparative performance was assessed on a held-out test set using both threshold-independent metrics (ROC-AUC and AP) and threshold-dependent classification outcomes (precision, recall, F1 derived from argmax predictions). The results demonstrate that both models provide strong discriminatory power for most anomaly categories, while XGBoost offers a substantial improvement in alert reliability and class separation in the more ambiguous risk type represented by patterned amounts.

#### 6.1. Model Comparison at Class Level: Separability and Alert Quality

Figure 11 summarizes the class-wise ranking and precision–recall performance for logistic regression and XGBoost on the test set. Both models achieve consistently high ROC-AUC values across the anomaly taxonomy, indicating strong separability for most classes. XGBoost shows a modest improvement for Regular Billing (AUC 0.965 vs. 0.954), confirming robust discrimination between normal and risk-relevant transactions. For Working-time anomaly, Edge of month, and Missing VAT, both models reach near-perfect performance (AUC and AP close to 1.00), which is consistent with these anomaly types being characterized by highly distinctive temporal and compliance indicators.



**Figure 11.** Per-class ROC-AUC and average precision (AP) for logistic regression vs. XGBoost.

The most remarkable difference is observed for the Patterned Amount class. While XGBoost provides a moderate improvement in ROC-AUC (from 0.899 to 0.922), the practical gain is substantially larger in precision–recall terms: Average Precision increases from 0.511 (logistic regression) to 0.827 (XGBoost). This improvement is particularly relevant for audit deployment because AP is sensitive to class imbalance and better reflects the expected alert quality, meaning that XGBoost can identify patterned-amount anomalies with considerably lower noise and therefore a more sustainable investigation workload. A similar, though smaller, improvement is observed for Shipping anomaly (AUC 0.980 vs. 0.969, AP 0.952 vs. 0.896), suggesting that the non-linear structure of XGBoost captures interaction effects that may not be adequately represented by a linear decision boundary.

The results shown in Figure 11 indicate that the added value of XGBoost is concentrated in the more ambiguous anomaly types, where risk signals are less deterministic and likely depend on non-linear combinations of operational and financial attributes. Ambiguous anomaly types in this context refer to transactions that simultaneously exhibit characteristics associated with more than one defined risk class. At the same time,

both models perform similarly and near-perfectly for anomaly classes driven by strong rule-like indicators.

### 6.2. Macro-Averaged Performance: Implications for Audit Operations

Beyond class-level metrics, the macro-averaged evaluation provides a balanced view that prevents the majority class from dominating the conclusions. The macro-averaged results presented in Table 9 in are based on argmax predictions, showing that XGBoost increases macro precision substantially (0.991 vs. 0.838 for logistic regression), while maintaining comparable macro recall (0.944 vs. 0.947). Consequently, the macro F1-score improves from 0.863 to 0.965. From an audit perspective, this pattern is particularly desirable because it indicates that XGBoost maintains sensitivity to anomalies while greatly reducing false positives across classes, thereby improving triage efficiency and increasing trust in the alerting mechanism.

**Table 9.** Macro-Averaged Model comparison—Loge vs. XGBoost classifier performance on Test data.

| Model               | Precision | Recall | F1    |
|---------------------|-----------|--------|-------|
| Logistic Regression | 0.838     | 0.947  | 0.863 |
| XGBoost             | 0.991     | 0.944  | 0.965 |

The improvement in macro precision suggests that XGBoost is better suited for continuous auditing contexts where human follow-up is costly and investigation resources are limited. Importantly, the similar recall values imply that this gain does not come at the expense of missing relevant anomalies, which would otherwise increase residual risk.

Although XGBoost achieved very high accuracy, the audit value of the model depends on the interpretation of classification errors. False positives increase audit workload because they require additional follow-up on transactions that may not represent actual exceptions. False negatives are more critical from an ERM perspective, as they may allow risky transactions, control failures, or potential fraud indicators to remain undetected. Therefore, model evaluation should be aligned with audit capacity and organizational risk appetite. Precision is relevant for controlling investigation costs, while recall is essential for reducing residual detection risk. For this reason, precision–recall analysis is particularly important in imbalanced audit datasets, where high overall accuracy may conceal weaker performance on rare but high-risk classes. An additional limitation concerns potential misclassification of commercially justified transactions as anomalies. Legitimate market fluctuations, such as seasonal demand spikes and promotional billing cycles, may generate unusual transaction concentrations that resemble audit anomalies but do not necessarily indicate fraud or control failure. The model outputs should therefore be interpreted as risk-based alerts supporting audit triage and professional judgment, rather than as definitive evidence of fraud.

### 6.3. Generalization and Robustness: Train-Test Consistency

The train vs. test macro-averaged results indicate stable generalization. Logistic regression remains consistent between train and test (precision 0.843 vs. 0.838, recall 0.953 vs. 0.947, F1 0.868 vs. 0.863). XGBoost exhibits similarly slight differences (precision 0.996 vs. 0.991, recall 0.951 vs. 0.944, F1 0.971 vs. 0.965). These small gaps provide evidence that the stronger performance of XGBoost is not primarily due to overfitting, but rather to improved modeling of non-linear relationships present in ERP billing behavior.

#### 6.4. Added Value to Internal Audit and ERM

This study contributes value in three complementary ways. First, it demonstrates that high-performing anomaly classification can be achieved using a limited set of audit-motivated indicators extracted from ERP billing data, thereby supporting the feasibility of scalable transaction monitoring. Second, it provides an evidence-based comparison between an interpretable linear baseline and a more expressive non-linear model, helping audit functions to choose a method aligned with their governance maturity and operational constraints. Third, by evaluating both ROC-AUC and PR-based metrics, the study emphasizes a realistic adoption perspective, where alert reliability and workload implications are as important as overall accuracy.

In practical deployment terms, logistic regression offers a transparent baseline with strong recall and intuitive coefficient-based interpretability, making it suitable for early-stage governance and stakeholder alignment. XGBoost, in turn, demonstrates markedly improved precision and overall F1 performance, particularly in ambiguous anomaly classes.

### 7. Conclusions and Future Work

This paper examined the applicability of supervised machine learning models for continuous auditing of ERP billing transactions, focusing on the early detection and classification of risk-relevant anomalies.

The audit engineered variables are a bridge between ML and audit for defining risky transactions, aligning with audit considerations, wider ERM approach, and the organization's maturity. The experimental results confirm that both multinomial logistic regression and XGBoost deliver strong predictive capability for several high-signal anomaly types, including working-time deviations, end-of-month patterns, and VAT-related issues.

However, the comparative evaluation highlights that XGBoost provides superior operational performance, primarily by improving precision and macro F1-score while maintaining high recall. This makes the XGBoost model particularly suitable for continuous monitoring scenarios where investigation capacity is constrained, and excessive false positives create significant delays in follow-up actions and limit the added value of automated flagging.

At the same time, logistic regression remains a valuable benchmark due to its transparency and easier governance. Its stable performance and interpretability make it appropriate for audit teams that prioritize explainability and policy traceability or for audit teams that are at an early stage of adopting AI-supported monitoring. The study demonstrates that audit-oriented machine learning can generate actionable, explainable signals from ERP data and thereby strengthen both internal audit planning and enterprise risk management by enabling earlier detection, improved prioritization, and more systematic coverage than traditional sampling-based approaches.

This study's scientific contribution is related to extending existing research by explicitly linking machine learning-based anomaly detection with enterprise risk management and continuous auditing frameworks, thereby positioning data-driven models within a broader governance and assurance context. The findings demonstrate how machine learning outputs can support audit decision-making, improve transaction monitoring, and reduce detection risk in real-world environments.

Future research can extend this work in several directions. First, it can broader validation across different organizations, such as manufacturing companies with high-volume purchase order processing and financial services firms with complex multi-entity billing structures. Additionally, larger fiscal periods and different ERP landscapes would strengthen external validity and allow assessment of model transferability under changing business rules and master data characteristics. Second, incorporating more contextual

features (e.g., customer segmentation, pricing procedure, authorization traces, or document flow linkages) could further reduce ambiguity in classes such as patterned amounts. Third, explainability and model control mechanisms should be further developed to support audit defensibility, for example, by integrating SHAP-based local explanations per flagged document. From a control perspective model documentation, data integrity validation and other controls in the application layer need to be considered. Finally, cost-sensitive optimization and thresholding strategies aligned with audit investigation costs and expected loss models would enable decision-theoretic alerting, improving the practical alignment between model outputs and risk-based audit response.

**Author Contributions:** Conceptualization, R.D., A.D., P.G. and M.L.; methodology, R.D. and A.D.; software, R.D. and A.D.; validation, R.D. and A.D.; formal analysis, R.D., A.D., P.G. and M.L.; investigation, R.D. and A.D.; resources, R.D. and A.D.; data curation, R.D. and A.D.; writing—original draft preparation, R.D. and A.D.; writing—review and editing, P.G. and M.L.; visualization, R.D. and A.D.; supervision, P.G. and M.L.; project administration, P.G. and M.L.; funding acquisition, P.G. and M.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** The research work presented in the paper is funded by the European Union-NextGenerationEU Via the National Recovery and Resilience Plan of the Republic of Bulgaria under project BG-RRP-2.004-0005 Improving the research capacity and quality to achieve international recognition and resilience of TU-Sofia (IDEAS).

**Data Availability Statement:** Restrictions apply to the datasets. The data presented in this study are not publicly available due to privacy and data protection regulations (GDPR). The data that support the findings of this study are available from the corresponding author upon reasonable request.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

|      |                                         |
|------|-----------------------------------------|
| AI   | Artificial Intelligence                 |
| ML   | Machine Learning                        |
| ERM  | Enterprise Risk Management              |
| ERP  | Enterprise Resource Planning            |
| CA   | Continuous Auditing                     |
| CM   | Continuous Monitoring                   |
| IDEA | Audit Sampling Software                 |
| ISA  | International Auditing Standards        |
| VAT  | Value Added Tax                         |
| PR   | Precision Recall                        |
| ROC  | Receiver Operating Characteristic Curve |
| AUC  | Area Under the Curve                    |

## References

- Ahmad, Maria, Grazia Marchesano, Vakentina Popolo, Roberto Revetria, and Anastasiia Rozhok. 2025. Development of a European Sustainability Reporting Standards Compliant Sustainability Assessment Framework for Manufacturing Organizations Using Analytic Hierarchy Process. *Sustainability* 17: 4772. [\[CrossRef\]](#)
- Akbaş, M. Çigdem. 2024. Measuring the Impact of Enterprise Risk Management on Performance, Value, and Risk Indicators of Borsa Istanbul XBANK Companies with Data Mining Prediction Models. *Humanities and Social Sciences Communications* 11: 1350. [\[CrossRef\]](#)
- Aladwey, Laila Mohamed Alshawadfy, and Samar El Sayad. 2024. Auditors' Perceptions of the Triggers and Obstacles of Continuous Auditing and Its Impact on Auditor Independence: Insights from Egypt. *Journal of Risk and Financial Management* 17: 578. [\[CrossRef\]](#)

- Alles, Michael, Gerard Brennan, Alexander Kogan, and Miklos A. Vasarhelyi. 2006. Continuous Monitoring of Business Controls: A Pilot Implementation of a Continuous Auditing System at Siemens. *International Journal of Accounting Information Systems* 7: 137–61. [\[CrossRef\]](#)
- Bakos, Levente, and Danut D. Dumitraşcu. 2021. Decentralized Enterprise Risk Management Issues under Rapidly Changing Environments. *Risks* 9: 165. [\[CrossRef\]](#)
- Bakumenko, Alexander, and Ahmed Elragal. 2022. Detecting Anomalies in Financial Data Using Machine Learning Algorithms. *Systems* 10: 130. [\[CrossRef\]](#)
- Berisha, Blend, Endrit Mëziu, and Isak Shabani. 2022. Big Data Analytics in Cloud Computing: An Overview. *Journal of Cloud Computing: Advances, Systems and Applications* 11: 75. [\[CrossRef\]](#)
- Best, Petar. J., Pall Rikhardsson, and Mark Toleman. 2009. Continuous Fraud Detection in Enterprise Systems through Audit Trail Analysis. *Journal of Digital Forensics, Security and Law* 4: 39–61. [\[CrossRef\]](#)
- Brandão, Andre, and Petia Georgieva. 2020. Log Files Analysis for Network Intrusion Detection. Paper presented at the 10th IEEE International Conference on Intelligent Systems (IS'20), Varna, Bulgaria, August 28–30; pp. 328–33. [\[CrossRef\]](#)
- Brown, Carol, Jeffrey Wong, and Amelia Baldwin. 2007. Research Streams in Continuous Audit: A Review and Analysis of the Continuous Auditing Literature. *Journal of Emerging Technologies in Accounting* 4: 1–14. [\[CrossRef\]](#)
- Bueno, Thales M., and Renan G. Canaan. 2024. The Brussels Effect in Brazil: Analysing the Impact of the EU Digital Services Act on the Discussion Surrounding the Fake News Bill. *Telecommunications Policy* 48: 102757. [\[CrossRef\]](#)
- Committee of Sponsoring Organizations of the Treadway Commission (COSO). 2018. Enterprise Risk Management—Integrating with Strategy and Performance. Available online: [https://www.coso.org/\\_files/ugd/3059fc\\_671ed4466c0e423b93a9ef3d2e30b786.pdf](https://www.coso.org/_files/ugd/3059fc_671ed4466c0e423b93a9ef3d2e30b786.pdf) (accessed on 27 November 2025).
- Davidson, Bruce I., Naman K. Desai, and Gregory J. Gerard. 2013. The Effect of Continuous Auditing on the Relationship between Internal Audit Sourcing and the External Auditor's Reliance on the Internal Audit Function. *Journal of Information Systems* 27: 41–59. [\[CrossRef\]](#)
- Di Palma, Gianmarco, Roberto Scendoni, Vittoradolfo Tambone, Rossana Alloni, and Francesco De Micco. 2025. Integrating Enterprise Risk Management to Address AI-Related Risks in Healthcare: Strategies for Effective Risk Mitigation and Implementation. *Journal of Healthcare Risk Management* 44: 25–33. [\[CrossRef\]](#)
- Duhova, Reneta, Angel Duhov, Milena Lazarova, and Petia Georgieva. 2025. Continuous Auditing with Machine Learning: Early Flagging of Risky Billing Documents from SAP. Paper presented at the 13th International Scientific Conference on Computer Science (COMSCI 2025), Sozopol, Bulgaria, September 13–15. [\[CrossRef\]](#)
- Grabski, Severin. V., Stewart A. Leech, and Pamela J. Schmidt. 2011. A Review of ERP Research: A Future Agenda for Accounting Information Systems. *Journal of Information Systems* 25: 37–78. [\[CrossRef\]](#)
- Gurtu, Amulya, and Jestin Johny. 2021. Supply Chain Risk Management: Literature Review. *Risks* 9: 16. [\[CrossRef\]](#)
- Gündoğdu, Serdar. 2023. Efficient Prediction of Early-Stage Diabetes Using XGBoost Classifier with Random Forest Feature Selection Technique. *Multimedia Tools and Applications* 82: 44531–53. [\[CrossRef\]](#) [\[PubMed\]](#)
- International Auditing and Assurance Standards Board (IAASB). 2009a. *ISA 200: Overall Objectives of the Independent Auditor and the Conduct of an Audit in Accordance with International Standards on Auditing*. New York: International Federation of Accountants.
- International Auditing and Assurance Standards Board (IAASB). 2009b. *ISA 330: The Auditor's Responses to Assessed Risks*. New York: International Federation of Accountants.
- International Auditing and Assurance Standards Board (IAASB). 2009c. *ISA 500: Audit Evidence*. New York: International Federation of Accountants.
- International Auditing and Assurance Standards Board (IAASB). 2019. *ISA 315 (Revised 2019): Identifying and Assessing the Risks of Material Misstatement*. New York: International Federation of Accountants.
- International Organization for Standardization. 2018. *Risk Management 2018—Guidelines*. ISO 31000:2018. Geneva: ISO.
- ISACA. 2022. *Audit Practitioner's Guide to Machine Learning, Part 1: Technology*. Schaumburg: ISACA. Available online: <https://www.isaca.org> (accessed on 28 November 2025).
- Ispas, Lucian, Costel Mironeasa, and Alessandro Silvestri. 2023. Risk-Based Approach in the Implementation of Integrated Management Systems: A Systematic Literature Review. *Sustainability* 15: 10251. [\[CrossRef\]](#)
- Jofre, Maria, and Richard Gerlach. 2018. Fighting Accounting Fraud through Forensic Data Analytics. *arXiv*, arXiv:1805.02840. [\[CrossRef\]](#)
- Knoblauch, Dorian, and Jurgen Großmann. 2023. Towards a Risk-Based Continuous Auditing-Based Certification for Machine Learning. *The Review of Socionetwork Strategies* 17: 255–73. [\[CrossRef\]](#)
- Knowles, Brian, and John T. Richards. 2021. The Sanction of Authority: Promoting Public Trust in AI. Paper presented at the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT'21), Virtual Event, Canada, March 3–10; pp. 262–71. [\[CrossRef\]](#)
- Kumar, Sonjai. 2022. Risk Management Framework. Available online: <https://ssrn.com/abstract=4141546> (accessed on 18 January 2026).

- Langley, Pat. 2011. The Changing Science of Machine Learning. *Machine Learning* 82: 275–79. [CrossRef]
- Le, Vinh Quang, Nguyen Ngoc-Long, and Pham Xuan Giang. 2024. Enterprise Risk Management and Firm Performance: Exploring the Roles of Knowledge, Technology, and Supply Chain. *Problems and Perspectives in Management* 22: 150–64. [CrossRef]
- Li, Nan, Lianbo Ma, Guo Yu, Bing Xue, Mengjie Zhang, and Yaochu Jin. 2023. Survey on Evolutionary Deep Learning: Principles, Algorithms, Applications, and Open Issues. *ACM Computing Surveys* 56: 1–36. [CrossRef]
- Locurcio, Marco, Francesco Tajani, Pierlugi Morano, Debora Anelli, and Benedetto Mangianelli. 2021. Credit Risk Management of Property Investments through Multi-Criteria Indicators. *Risks* 9: 106. [CrossRef]
- Murikah, Wilberforce, Jeff K. Nthenge, and Faith M. Musyoka. 2024. Bias and Ethics of AI Systems Applied in Auditing—A Systematic Review. *Scientific African* 25: e02281. [CrossRef]
- Polizzi, Ssalvatore, and Enzo Scannella. 2023. Continuous Auditing in Public Sector and Central Banks: A Framework to Tackle Implementation Challenges. *Journal of Financial Regulation and Compliance* 31: 40–59. [CrossRef]
- Samuel, Arthur L. 1959. Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development* 3: 210–29. [CrossRef]
- SAP SE. 2024. *What Is SAP ERP? Key Features & Benefits*. Walldorf: SAP SE. Available online: <https://www.sap.com/products/erp/what-is-sap-erp.html> (accessed on 5 December 2025).
- Shapovalova, Alla, Olena Kuzmenko, Oleg Polishchuk, Tetyana Larikova, and Zorana Myronchuk. 2023. Modernization of the National Accounting and Auditing System Using Digital Transformation Tools. *Financial and Credit Activity: Problems of Theory and Practice* 4: 33–53. [CrossRef]
- Shin, Il H., Myuong G. Lee, and Woojin Park. 2013. Implementation of the Continuous Auditing System in the ERP-Based Environment. *Managerial Auditing Journal* 28: 592–627. [CrossRef]
- Tarafder, Agnidipto, and Aniruddh Vadlamani. 2024. Will the EU AI Regulations Give Rise to Another ‘Brussels Effect’? Lessons from the GDPR. *Journal of Development Policy and Practice* 10: 45–60. [CrossRef]
- Țircovnicu, Georgiana I., and Camelia D. Hățegan. 2023. Integration of Artificial Intelligence in the Risk Management Process: An Analysis of Opportunities and Challenges. *Journal of Financial Studies* 8: 198–214. [CrossRef]
- Xiao, Wensheng, Yu Tang, Bright Obuobi, Shaojian Qu, Minglan Yuan, and Decai Tang. 2023. The Influence of Rule of Law on Government’s Sustainable Economic Policy. *Sustainability* 15: 11690. [CrossRef]
- Yuan, Tingyu, Xi Zhang, and Xuanjing Chen. 2025. Machine Learning Based Enterprise Financial Audit Framework and High Risk Identification. *arXiv*, arXiv:2507.06266. [CrossRef]
- Zeng, S., and H. Li. 2025. Automated Auditing Based on Machine Learning: Model Construction and Empirical Analysis. *Computer Fraud & Security* 2025: 544–52. [CrossRef]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.