Journal of Data Analytics and Intelligent Decision-Making
Journal homepage: jdaid.qom.ac.ir
University of Qom

Improving Multi-Label Business Text Classification with Imbalanced Data: Adjusted BCE Weighting and Threshold Optimization for Rare Labels in BERT Models

Sina Hassani [a], Haleh Homayouni [a]*, Kimia Bazargan Lari [a], Danial Soleimani [a]

a. Apadana Institute, Shiraz, Iran,.

ARTICLEINFO

Keywords:
Multi-label Classification; Weighted Loss Function; Business Text Classification; Imbalanced Learning

ABSTRACT

Multi-label classification of business texts in the presence of imbalanced label distributions remains a significant challenge in Natural Language Processing. Tail labels, which are associated with very few training samples, typically exhibit weak predictive performance even when advanced transformer-based models such as BERT are employed. This limitation hinders the reliable identification of rare but potentially valuable business opportunities within large-scale textual data. The present study aims to enhance tail-label performance by introducing an adjusted weighting strategy into the Binary Cross-Entropy (BCE) loss function. The proposed approach consists of two main components. First, a label-specific weight is calculated as the ratio of negative to positive samples for each label and then constrained within a predefined range to prevent excessive dominance of either frequent or rare labels. Second, an optimal decision threshold is determined through grid search over the interval [0.1, 0.9], enabling improved balance between precision and recall across labels. Experiments are conducted on an English multi-label dataset containing 1,000 samples and 20 imbalanced labels, with label frequencies varying from 180 to 5 instances. The data are split into 80% training and 20% testing sets. Results show that the weighted BERT model achieves a Hamming accuracy of 0.623, a macro-F1 score of 0.091, and a tail-label F1 score of 0.025. Notably, using only one twenty-eighth of the baseline dataset size, the model retains approximately 70% of baseline accuracy while improving tail-label performance compared to the unweighted setting. The method offers a practical, computationally efficient solution for data-scarce and resource-constrained environments.

* Corresponding author.
E-mail addresses: haleh.homayouni@gmail.com (H. Homayouni).
Received 15 Jan 2026; Received in revised form 17 Mar 2026; Accepted 29 Mar 2026
Available online 30 Mar 2026

rticle: Hassani, et al. (2026). Improving Multi-Label Business Text Classification with Imbalanced Data: Adjusted BCE Weighting and Threshold Optimization for Rare Labels in BERT Models.
Journal of Data Analytics and Intelligent Decision-Making, 2(1),74-90.
https://doi.org/10.22091/10.22091/jdaid.2026.15478.1039

1. Introduction

In recent years, organizations have increasingly relied on enterprise information systems, such as Enterprise Resource Planning (ERP), Customer Relationship Management (CRM), and workflow management systems. These systems not only support the execution of business processes but also continuously record substantial volumes of event data. Such data, stored in the form of event logs, provide detailed traces of how processes are actually executed, including the sequence of activities, resource allocations, and the flow of cases over time. Access to these data enables organizations and researchers to analyze processes beyond predefined designs and instead examine them based on their real operational behavior.

Within this context, process mining has emerged as a data-driven approach that bridges Business Process Management (BPM) and data science, allowing the extraction of meaningful knowledge from event logs (van der Aalst, 2016). Unlike traditional BPM approaches, which primarily rely on expert interviews, analytical workshops, or manual modeling, process mining utilizes data generated during the actual execution of processes to reconstruct models grounded in empirical evidence. This capability has gradually established process mining as a significant research area in academia and a practical tool in industry (Augusto et al., 2019; van der Aalst et al., 2011). Numerous studies have also highlighted the growing application of process mining in domains such as healthcare, financial services, manufacturing, and smart service systems (Marrella, 2019; Rebuge & Ferreira, 2012; Rojas et al., 2016).

Generally, process mining encompasses three main areas: process discovery, conformance checking, and process enhancement (van der Aalst, 2016). Among these, process discovery plays a central role as it often represents the starting point for subsequent analyses. Process discovery aims to automatically extract a process model from event logs without relying on prior knowledge of the process structure. The resulting model serves as the foundation for further analyses such as conformance checking and performance evaluation; therefore, the quality of the discovered model directly affects the validity of these subsequent analyses. Consequently, evaluation criteria, such as fitness, precision, generalization, and simplicity, have received significant attention in the literature (Adriansyah et al., 2011; Buijs et al., 2014).

Since the introduction of the Alpha algorithm as one of the earliest process discovery techniques (van der Aalst et al., 2004), numerous studies have attempted to address its limitations, including sensitivity to noise, structural complexity, and scalability issues. Heuristic-based methods, such as the Heuristics Miner, sought to improve robustness against noisy data by exploiting the frequency of relationships between events (Weijters & van der Aalst, 2003). Subsequently, inductive approaches, such as the Inductive Miner, provided guarantees regarding the structuredness of discovered models, representing a significant step toward improving the practical usability of process discovery results (Leemans et al., 2013; Leemans et al., 2014). More advanced techniques have since been proposed, including probabilistic modeling, decomposition strategies, and metaheuristic methods aimed at improving stability and scalability (Augusto et al., 2019; Carmona et al., 2022).

Alongside these developments, increasing attention has been directed toward integrating machine learning and deep learning techniques into process discovery. These approaches, particularly when dealing with large, variable, or incomplete logs, have demonstrated the ability to identify more complex control-flow patterns (Camargo et al., 2019; Pasquadibisceglie et al., 2020; Tax & Van der Aalst, 2016). However, the adoption of these methods not only addresses existing challenges but also introduces new concerns, including reduced model interpretability and a lack of consensus on appropriate evaluation frameworks.

Despite these advancements, the existing review literature on process discovery remains limited in several important respects. First, many previous reviews concentrate on specific families of algorithms (e.g., heuristic-based or inductive approaches) rather than providing a

comprehensive synthesis of the entire methodological landscape. Second, prior reviews often emphasize technical algorithmic descriptions while offering limited discussion of evaluation criteria, scalability considerations, and practical applicability in real organizational environments. Third, emerging research directions—such as object-centric process discovery, hybrid AI-driven approaches, and data-intensive discovery techniques—are often examined separately rather than within an integrated analytical framework. As a result, existing review literature provides valuable but fragmented insights into the evolution and comparative characteristics of process discovery methods.

These limitations underscore the need for a comprehensive and systematically structured review that integrates methodological developments, evaluation practices, and emerging research directions within a unified perspective. Such a synthesis is particularly important for both researchers and practitioners, as the growing diversity of algorithms and modeling approaches makes it increasingly difficult to understand their relative strengths, limitations, and applicability in real-world contexts.

Accordingly, this paper presents a systematic review of process discovery methods in process mining. Unlike previous reviews that focus on narrow subsets of techniques or specific application domains, this study provides an integrated analysis of process discovery methods by combining three complementary perspectives: the methodological classification of existing approaches, the examination of evaluation practices used to assess discovered models, and the identification of emerging research trends and unresolved challenges.

The main contributions of this study are threefold. First, the paper synthesizes fragmented research on process discovery by offering a structured classification of dominant methodological approaches. Second, it analyzes the evaluation criteria and experimental practices used in the literature, providing insights into how the quality of discovered models is assessed. Third, it identifies key research gaps and challenges—including issues related to scalability, interpretability, and integration with advanced data-driven techniques—that shape future research directions in process mining. To ensure methodological rigor and transparency, this review follows established guidelines for systematic literature studies in software engineering and information systems (Kitchenham, 2004; Petersen et al., 2015), enabling a clear and reproducible process for study selection, analysis, and synthesis.

2. Theoretical Foundations and Related Work

Process mining, as an emerging and interdisciplinary field within data science, plays an important role in improving operational efficiency and supporting organizational decision-making. By leveraging data recorded in information systems, this field enables the modeling, analysis, and improvement of real organizational processes (van der Aalst, 2011). Process mining can be considered a step beyond traditional data mining, as its focus extends past simple pattern discovery to the reconstruction of the behavioral and control-flow structures of processes. Through the extraction of knowledge from event logs, process mining techniques enable the automatic discovery of process models, the assessment of conformance between actual behavior and existing models, and the enhancement and enrichment of implemented organizational process models (van der Aalst, 2016). These capabilities have made process mining an effective tool in business process management, software system analysis, and the optimization of organizational operations, attracting significant attention from both researchers and practitioners. Accordingly, the research background related to the present study is summarized in Table 1.

To identify prior research related to process discovery methods in process mining, a systematic and multi-stage search strategy was adopted. To achieve this, several major international databases, including Google Scholar, IEEE Xplore, Springer, and Scopus

(Elsevier), as well as the Iranian academic database Noormags, were examined. The search was conducted focusing on titles, abstracts, and keywords, using combinations of the main terms "process mining," "process discovery," and "algorithm," along with the logical operators AND and OR, in order to achieve comprehensive coverage of relevant studies.

The search covered the period from 2000 to 2023, a time-frame that encompasses both the early emergence of process mining and the recent developments in the field. Given the large volume of retrieved studies and the considerable overlap among some of them, particular attention was paid during the screening phase to the selection of the most relevant and influential works. Ultimately, a summary of the identified studies is presented in Table 1, providing a clear and traceable overview of the research trends in this area.

Table 1
Review of Related Studies

Article/Study Title	Research Method	Research Objective	Key Findings	Citation
A Survey of Process Mining for Customer Management	Literature review	Examining the application of process discovery in customer experience analysis	Major process discovery algorithms such as Heuristic, Alpha, and Inductive Miner are widely used in customer process analysis.	Dioses & Cordova (2025)
On Process Discovery Experimentation: Addressing the Challenges	Applied/experimental study	Investigating practical challenges in experimenting with process discovery algorithms	The appropriate selection of evaluation metrics and experimental datasets is one of the most critical challenges in assessing process discovery methods.	Rehse (2024)
A Survey on Concept Drift in Process Mining	Literature review	Investigating the phenomenon of concept drift in process mining and its impact on process discovery	Traditional models assume process stability and are not well suited for dynamic or evolving processes.	Vecino Sato et al. (2021)
Event Log Generation: An Industry Perspective	Empirical study	Examining challenges related to generating event logs and their impact on process mining	Key challenges include data quality and information integration, which directly affect process discovery outcomes.	Kampik & Weske (2022)
Comparing Ordering Strategies for Process Discovery	Algorithmic study	Investigating the role of event ordering in the quality of discovered models	Proper ordering strategies can improve model accuracy and reduce computational time.	Huang & van der Aalst (2023)
Log Skeletons: A Classification Approach to Process Discovery	Conceptual research	Evaluating the effectiveness of log skeleton models in process discovery	Log skeleton models outperform some fully automated algorithms in	Verbeek & de Carvalho (2018)

			behavior classification tasks.	
Process Mining Techniques and Applications: A Systematic Mapping Study	Systematic mapping study	Identifying active research topics in process mining	Process discovery is one of the core research themes and has broad applications across domains.	Dos Santos Garcia et al. (2019)
Process Mining in Organizational Environments: A Systematic Review	Systematic review	Examining research needs and trends in organizational process mining	Early research focused on process modeling, while more recent studies emphasize real organizational applications.	Norouzi & Yelveh (2025)
Integration of Machine Learning in Process Discovery	Literature review	Investigating the integration of machine learning with process discovery methods	Integrating machine learning can improve the quality of discovered process models but requires high-quality data.	Lee & Kim (2023)
Robust Process Mining with Guarantees	Review book/conceptual framework	Presenting advanced frameworks for discovering and evaluating process models	Introduces the Inductive Miner framework and discusses quality guarantees in discovered models.	Leemans (2022)
Customer-Centric Process Discovery Approaches	Article review	Reviewing specialized process discovery methods for customer journey analysis	Algorithms must capture stage-based and decision-oriented states within event logs.	Dioses & Cordova (2025)
Explainability in Process Mining: A Survey	Literature review	Examining explainability methods in process mining	Explainable approaches help interpret models more effectively and highlight interpretability challenges in discovered processes.	Mehdiyev et al. (2023)
Process Mining for Unstructured Data: Challenges and Solutions	Literature review	Investigating challenges of unstructured data in process mining	Unstructured data increases analytical complexity and requires advanced preprocessing techniques.	Koschmider (2024)
Scalable Process Discovery and Conformance Checking	Framework-based research	Analyzing scalability challenges in process discovery and conformance checking	Proposes frameworks to improve scalability for large event logs.	van der Aalst (2016)
Robotic Process Automation Using Process Mining	Systematic review	Examining the role of process mining in supporting RPA	Process mining can identify automatable process structures but requires	El-Gharib (2023)

		through process discovery	careful event log preprocessing.	
Data Quality in Process Mining: A Systematic Review	Systematic review (Persian)	Investigating data quality challenges and their impact on process discovery	Data quality is identified as one of the most critical challenges for applying process discovery in real environments.	Salehi et al. (2023)

3. Methodology

This study adopts a Systematic Literature Review (SLR) approach to provide a comprehensive and transparent synthesis of research on process discovery methods in process mining. The review process follows the PRISMA guidelines and was conducted through a structured sequence of activities, including literature search, screening, eligibility assessment, quality evaluation, and final analysis.

The literature search was performed across several major academic databases, including Google Scholar, IEEE Xplore, Springer, Scopus (Elsevier), and the Iranian scientific database Noormags. These databases were selected to ensure broad coverage of peer-reviewed journal articles and conference papers in the fields of process mining, information systems, and computer science. The search strategy was designed iteratively and employed combinations of the keywords “process mining,” “process discovery,” and “algorithm” using logical operators. The search was applied to titles, abstracts, and keywords. Both English and Persian publications were considered, and the search covered studies published between 2000 and 2023. This process initially identified 4,940 records.

All retrieved records were merged into a single dataset, and duplicate entries were removed using a combination of automated filtering and manual verification. The remaining studies were then screened based on their titles and abstracts. During this stage, several exclusion criteria were applied. Studies were excluded if they were unrelated to process mining, did not focus on process discovery, represented non-scientific materials such as tutorials, editorials, or reports, or lacked a clear methodological contribution.

The full texts of the remaining articles were then examined based on predefined inclusion and exclusion criteria. To be included in the review, studies had to explicitly address process discovery within process mining, propose, evaluate, or compare a process discovery method or algorithm, be published as a peer-reviewed journal or conference paper, and provide sufficient methodological detail to allow interpretation and comparison. Studies were excluded if their primary focus was on other areas of process mining, such as conformance checking or process enhancement, if they only mentioned process discovery marginally, if they lacked methodological clarity, or if they represented preliminary or duplicated versions of previously published work.

Following the eligibility assessment, a quality evaluation process was conducted to ensure the methodological robustness of the selected studies. Each article was assessed according to several quality criteria, including the clarity of the proposed method or algorithm, the adequacy of the methodological description, the presence of experimental evaluation or empirical validation, and the overall relevance to process discovery research. Studies that did not meet the minimum quality standards were excluded from further analysis. After completing the screening and quality assessment stages, a total of 138 studies were retained for the final analysis, and the entire selection process was documented using a PRISMA flow diagram.

For the selected studies, a structured data extraction process was performed. Key information was collected from each article, including publication year, publication venue

(journal or conference), the type of process discovery method employed, the modeling techniques used, evaluation approaches, characteristics of the event logs or datasets, and application domains.

To analyze the collected data, descriptive statistical methods were used to examine publication trends from 2000 to 2023, the distribution across journals and conferences, and the evolution of methodological approaches. Time-series trends were further analyzed using ARIMA models, with model validity evaluated through ACF and PACF tests.

In addition to statistical analysis, an iterative thematic analysis was conducted to identify the main methodological patterns within the selected studies. During this process, the extracted methodological characteristics of the studies were repeatedly reviewed, compared, and grouped based on conceptual similarities. Through this iterative coding and refinement process, the studies were ultimately organized into four main categories of process discovery approaches: algorithmic and heuristic approaches, rule- and constraint-based approaches, probabilistic and sequence-based models, and machine learning and deep learning-based methods.

Throughout the entire review process, all stages—from search and screening to data extraction and analysis—were carefully documented and conducted using consistent criteria. Adherence to the PRISMA framework enhances the transparency, reproducibility, and reliability of the review. This methodology therefore provides a systematic and structured synthesis of existing knowledge on process discovery methods while enabling the identification of research trends, methodological developments, and remaining challenges in the field.

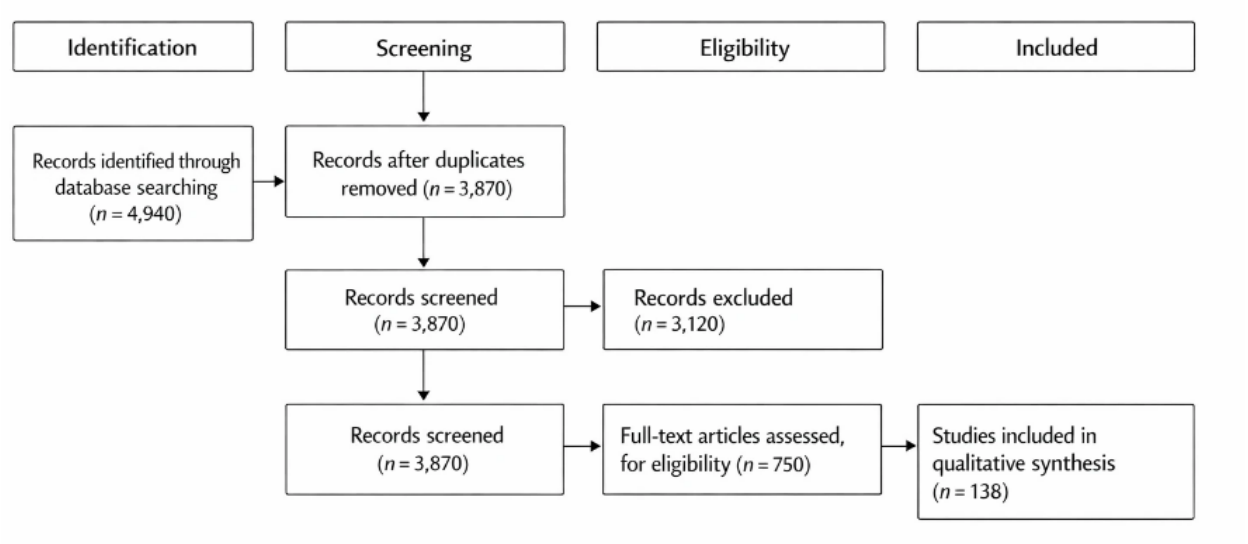

Figure 1
PRISMA Flow Diagram of the Study Selection Process

4. Results

This section presents and analyzes the findings derived from the systematic review of 138 selected articles. The analysis mainly focus on identifying the dominant approaches to process discovery, examining prevailing research trends, and extracting recurring challenges reported in the literature. Given the breadth of the topic and the heterogeneity of the studies in terms of objectives, datasets, and applied methodologies, the results are reported at a high-level analytical perspective to provide a coherent and generalizable overview of the state of research in this field.

5. Main Approaches to Process Discovery

The review of the selected studies indicates that process discovery methods can be categorized into four major approaches (Table 2). Among these, algorithmic and rule-based approaches are

considered the classical core of process mining. Due to their high interpretability, well-defined formal structure, and ability to support behavioral analysis of process models, these methods continue to hold a dominant position in the literature (Leemans et al., 2013; Van der Aalst et al., 2004). However, the sensitivity of these algorithms to noise and the complexity of real-life event logs has introduced limitations for their practical application.

In response to these challenges, heuristic-based methods have been developed, relying on the frequency and strength of relationships between events. These approaches offer greater flexibility when dealing with incomplete or noisy data and have been widely used in empirical studies analyzing real event logs (Weijters & Van der Aalst, 2003). Nevertheless, a reduction in the formal accuracy of resulting models in certain scenarios is considered one of the main limitations of this category of methods.

The third category includes evolutionary and optimization-based approaches, which are designed to simultaneously optimize multiple model quality criteria, such as fitness, simplicity, and generalization. These methods generally provide greater flexibility in discovering complex process structures. However, their high computational cost and long execution times have limited their widespread adoption, particularly in large-scale industrial environments (De Medeiros et al., 2007).

In recent years, data-driven and machine learning-based approaches have emerged as a growing trend in process discovery research. These methods—particularly sequence-based models and neural networks—have demonstrated a stronger capability in identifying complex control-flow patterns and handling large, dynamic event logs (Evermann & Tax, 2017; Tax & Van der Aalst, 2016). However, the findings suggest that the increased modeling power of these approaches is often accompanied by reduced interpretability and the absence of standardized evaluation frameworks, an issue that remains an open challenge in the literature.

Table 2
High-level Classification of Process Discovery Approaches (Aligned with "Approaches" in the Article Title)

Approach Category	Core Characteristics	Representative Methods	Sample References
Rule-based/Algorithmic	Formal guarantees, high interpretability, structured discovery	Alpha Miner, Inductive Miner	van der Aalst et al. (2004); Leemans et al. (2013)
Heuristic-based	Noise-tolerant, frequency-driven discovery	Heuristic Miner	Weijters & van der Aalst (2003)
Evolutionary/Meta-heuristic	Multi-objective optimization, flexible model search	Genetic Process Mining	de Medeiros et al. (2007)
Data-driven/ML-based	Scalability, adaptability, pattern learning	Clustering-based, Neural models	Evermann et al. (2017); Tax & Van der Aalst (2016)

6. Identified Research Trends

An examination of the temporal and thematic distribution of the selected studies indicates that research on process discovery has gradually shifted over the past two decades from a primary focus on classical algorithms toward the integration of diverse approaches and the utilization of artificial intelligence techniques and real-world data (Table 3). A high-level analysis of the findings reveals several recurring research trajectories, which are discussed below.

First, the growing tendency toward hybrid approaches emerges as one of the most prominent trends. In these studies, researchers attempt to combine traditional algorithmic structures with machine learning capabilities to enhance the flexibility and adaptability of

process models (Evermann & Tax, 2017; Van der Aalst, 2016). The primary objective is to improve the model's ability to handle noisy data while minimizing the interference caused by process complexity. These hybrid approaches represent a convergence between formal process modeling and data-driven learning, thereby opening new directions for future research.

A second observable trend is the increasing attention to noisy and non-ideal event logs, reflecting the research community's movement toward real-world and industrial applications of process mining. In contrast to earlier studies that often relied on simulated or controlled datasets, recent research places greater emphasis on complex real-life event logs (Mannhardt et al., 2016). Nevertheless, the use of large-scale industrial datasets remains relatively limited. This limitation is primarily associated with organizational confidentiality concerns and restricted access to operational data (Bose et al., 2012), a challenge that is particularly evident in sectors such as financial services and healthcare.

A third notable trend is the emergence of online process discovery as a developing research area. This line of research focuses on analyzing streaming event data in order to monitor and model processes in real time (Van Zelst et al., 2018). Despite the significant potential of this approach for supporting real-time decision-making, the findings indicate that current research remains at an early stage of development. Technical aspects, scalability issues, and evaluation methodologies still require further advancement.

Overall, the temporal and thematic analysis of the literature suggests a gradual transition from formal, algorithm-centric approaches toward data-driven and adaptive models. This shift has been accelerated by the increasing volume of organizational data and the growing complexity of business processes. While this new direction enhances the practical applicability of process discovery methods, it simultaneously raises new questions regarding interpretability, transparency, and model validity—issues that are discussed in greater detail in the challenges Section.

Table 3

Identified Research Trends in Process Discovery (Aligned with "Trends" in the Article Title)

Research Trend	Description	Evidence in Literature	Sample References
Shift towards hybrid approaches	Combination of classical miners with ML techniques	Increasing number of hybrid proposals after 2016	van der Aalst (2016); Evermann et al. (2017)
Increased focus on noisy real-life logs	Emphasis on robustness over formal optimality	Frequent use of heuristic and filtering mechanisms	Mannhardt et al. (2016)
Limited use of large-scale industrial logs	Predominance of benchmark or synthetic datasets	Recurrent limitation statements	Bose et al. (2012)
Emerging interest in online/streaming discovery	Early-stage research with limited adoption	Few but growing studies	van Zelst et al. (2018)

7. Common Challenges in the Process Discovery Literature

Despite the considerable diversity of approaches and the methodological advancements reported in the process discovery literature, the systematic review reveals that several recurring challenges remain unresolved. These challenges are largely interdisciplinary in nature and are not limited to a single category of methods; rather, they appear across most research streams in the field.

One of the most prominent challenges concerns the scalability of process discovery methods when dealing with large, complex, and heterogeneous event logs. This issue is

particularly evident in evolutionary and data-driven approaches, where increases in data volume directly led to higher computational costs and longer execution times (Bose et al., 2012; Tax & Van der Aalst, 2016). Although some studies have proposed techniques such as process decomposition, probabilistic modeling, and metaheuristic methods to mitigate this limitation, the evidence suggests that scalability remains one of the primary barriers to the widespread industrial adoption of these techniques.

Another major challenge is the decline in model interpretability, particularly in approaches based on machine learning and deep learning. While these methods demonstrate strong capabilities in identifying complex control-flow patterns and handling highly variable event logs, their structural complexity often makes it difficult for organizational analysts to understand process behavior and conduct causal analysis of the resulting models (Evermann & Tax, 2017). This issue creates a fundamental trade-off between model accuracy and simplicity: more complex models may achieve better predictive or fitting performance, yet they may be less suitable for managerial decision-making and practical adoption.

In addition, the absence of standardized evaluation and benchmarking frameworks has been identified as a structural challenge within the literature. Variations in evaluation metrics, datasets, and experimental scenarios make it difficult to conduct fair and systematic comparisons among different methods (Buijs et al., 2014; Van der Aalst, 2016). This situation not only affects the reproducibility of research results but also constitutes a significant barrier to the accumulation of knowledge and the transfer of research outcomes into practice.

Finally, the analysis of the reviewed studies indicates the existence of a significant gap between theoretical method development and the practical needs of organizations. Limited access to industrial datasets, confidentiality concerns, and the discrepancy between laboratory conditions and operational environments have caused many proposed methods to remain largely at the research stage. From this perspective, the practical adoption of process discovery techniques, particularly in engineering management and software systems, depends less on the introduction of new algorithms and more on model transparency, comparability, and alignment with real organizational challenges.

Table 4

Cross-Cutting Challenges Reported in Process Discovery Literature (Aligned with "Challenges" in the Article Title)

Challenge	Description	Affected Approaches	Sample References
Scalability	Performance degradation with large or complex logs	ML-based, Evolutionary	Bose et al. (2012); Tax & Van der Aalst (2016)
Interpretability	Reduced transparency in data-driven models	ML-based approaches	Evermann et al. (2017)
Lack of standardized evaluation	Inconsistent metrics and benchmarks	All approaches	Buijs et al. (2014)
Reproducibility	Limited access to datasets and implementations	All approaches	van der Aalst (2016)

8. Discussion and Conclusion

This systematic review was conducted to clarify the current state of research on process discovery within the broader field of process mining and to synthesize the methodological directions and challenges reported in the literature. By systematically analyzing 138 selected studies, this research provides a structured overview of the main methodological approaches, the evolution of research trends, and the practical challenges that influence the adoption of process discovery techniques in real-world environments. The results indicate that while

significant progress has been achieved in the development of new algorithms and modeling techniques, many studies increasingly focus on issues beyond pure algorithmic innovation, including scalability, evaluation practices, and the applicability of discovered models in complex organizational settings.

The findings show that research in process discovery has evolved through the coexistence of classical algorithmic approaches and more recent data-driven methods. Classical methods, including algorithmic and heuristic approaches, remain important due to their interpretability and structural transparency, making them suitable for organizational analysis and managerial decision making. In contrast, machine learning and deep learning–based methods have been introduced primarily to address challenges such as large-scale event logs, data noise, and the complexity of modern software-intensive processes. The increasing presence of hybrid approaches in the literature reflects an attempt to combine the interpretability of traditional models with the adaptability and scalability of data-driven techniques. This trend suggests a gradual transition in the field toward more integrated methodological frameworks.

Another important observation concerns the growing attention to noisy and real-world event logs. Many recent studies attempt to design discovery techniques that are capable of handling incomplete, inconsistent, or highly variable data. This shift indicates a broader recognition of the practical conditions under which organizational process data are generated. However, despite this progress, the analysis also reveals that a relatively limited number of studies rely on large-scale industrial datasets. This limitation highlights a persistent gap between academic research and the operational needs of real organizational environments, particularly in areas such as project management, enterprise systems, and complex software development processes.

The review also identifies several recurring challenges that continue to shape research in this domain. Among the most frequently reported issues are scalability limitations, reduced interpretability of complex models, and the absence of widely accepted evaluation frameworks. These challenges are not confined to specific algorithms but reflect broader methodological issues that affect the reliability and comparability of research results. From an engineering and systems management perspective, the lack of standardized evaluation approaches makes it difficult to compare different discovery techniques and to determine their suitability for specific organizational contexts.

Another key challenge concerns the balance between model accuracy and interpretability. While advanced machine learning–based methods often achieve higher predictive or discovery performance, their outputs may be difficult for managers and system analysts to interpret. This trade-off between analytical performance and explainability suggests that algorithmic improvements alone may not necessarily translate into improved decision support for organizations. As a result, the practical value of process discovery methods increasingly depends on their ability to produce understandable and actionable models rather than merely achieving higher computational performance.

Overall, the results of this systematic review indicate that the evolution of process discovery research reflects broader changes in organizational and technological environments. Modern organizations operate in highly dynamic and data-intensive contexts, which require analytical tools capable of handling complex, heterogeneous, and continuously evolving process data. The growing interest in hybrid, probabilistic, and machine learning–based approaches can therefore be interpreted as a response to these emerging requirements. Nevertheless, the available evidence suggests that the adoption of these advanced techniques in real organizational settings remains limited.

One of the main conclusions of this study is that the practical success of process discovery techniques depends not only on algorithmic accuracy but also on factors such as model

interpretability, transparency of assumptions, comparability of evaluation results, and the ability to integrate with existing organizational information systems. Methods that generate highly complex or opaque models may face barriers to adoption, particularly in environments where managerial understanding and explainability are essential for decision making.

Based on these observations, future research in process discovery should move beyond the exclusive development of increasingly sophisticated algorithms. Instead, greater emphasis should be placed on the alignment of methodological innovations with the needs of engineering management, software systems development, and organizational process improvement. In particular, future studies may benefit from focusing on explainable discovery models, standardized evaluation methodologies, broader use of real industrial datasets, and the integration of human and organizational factors into process mining research. Addressing these aspects can help reduce the gap between academic research and practical implementation, thereby strengthening the role of process mining as a reliable analytical tool for organizational process management and engineering applications.

References

Augusto, A., Conforti, R., Dumas, M., La Rosa, M., Maggi, F. M., Marrella, A., ... & Soo, A. (2018). Automated discovery of process models from event logs: Review and benchmark. *IEEE transactions on knowledge and data engineering*, *31*(4), 686-705.

Bose, R., van der Aalst, W. M. P., & Zomerdijk, B. (2012). Process mining in business environments: Challenges and opportunities. *Journal of Management Information Systems, 29*(4), 3-21. https://doi.org/10.2753/MIS0742-1222290401

Buijs, J. C. A. M., van Dongen, B. F., & van der Aalst, W. M. P. (2014). Quality dimensions in process discovery: The importance of fitness, precision, generalization and simplicity. *International Journal of Cooperative Information Systems, 23*(1), 1440001. https://doi.org/10.1142/S0218843014400017

Camargo, M., Dumas, M., & Rojas, O. G. (2019). Simod: a tool for automated discovery of business process simulation models. In *BPM (PhD/Demos)* (pp. 139-143).

Carmona, J., van Dongen, B., & Weidlich, M. (2022). Conformance checking: Foundations, milestones and challenges. In *Process mining handbook* (pp. 155-190). Springer International Publishing.

De Medeiros, A. A., Weijters, A. J. M. M., & Van der Aalst, W. M. P. (2007). Genetic process mining: An experimental evaluation. In *1st International Conference on Business Process Management* (pp. 1-16). Springer. https://doi.org/10.1007/978-3-540-75183-7_1

Dioses, J., & Cordova, L. (2025). A survey of process mining for customer management. *Engineering Proceedings, 83*(1), 7. https://doi.org/10.3390/26734591(83)7 (MDPI)

dos Santos Garcia, C., Meincheim, A., Junior, E. R. F., Dallagassa, M. R., Sato, D. M. V., Carvalho, D. R., ... & Scalabrin, E. E. (2019). Process mining techniques and applications–A systematic mapping study. *Expert Systems with Applications*, *133*, 260-295.

El-Gharib, N.M. (2023). Robotic process automation using process mining: A systematic literature review. *Decision Support Systems, 148*, 102229.

Evermann, J., & Tax, N. (2017). Process mining with machine learning: Challenges and opportunities. *Computers in Industry, 91*, 52-59. https://doi.org/10.1016/j.compind.2017.07.010

Huang, T.-H., & van der Aalst, W.M.P. (2023). Comparing ordering strategies for process discovery using synthesis rules. *arXiv*, *2301*.02182.

Kampik, T., & Weske, M. (2022). Event log generation: An industry perspective. *arXiv*, *2202*.02539.

Koschmider, A. (2024). Process mining for unstructured data: Challenges and solutions. *GI Proceedings*. (GI Download)

Lee & Kim (2023). Integration of machine learning in process discovery: Trends and techniques. *Artificial Intelligence Review*.

Leemans, S., Mans, R., & Van der Aalst, W. M. P. (2013). Discovering block-structured process models from event logs. In *9th International Conference on Business Process Management* (pp. 296-313). Springer. https://doi.org/10.1007/978-3-642-40176-1_20

Leemans, S. J. J. (2022). *Robust process mining with guarantees: Process discovery, conformance checking and enhancement*. Springer.

Mannhardt, F., et al. (2016). Mining noise-robust process models from event logs. *Information Systems, 61*, 43-57. https://doi.org/10.1016/j.is.2016.03.004

Mehdiyev, N., Majlatow, M., & Fettke, P. (2023). Interpretable and explainable machine learning methods for predictive process monitoring: A systematic literature review. *arXiv, 2312*.17584.

Norouzi, Y., & Yalveh, E. (2025). Process mining in organizational environments: A systematic literature review. *Academic Librarianship and Information Research*, *59*(1), 1-23. https://doi.org/10.22059/jlib.2025.388900.1769

Pasquadibisceglie, V., Appice, A., Castellano, G., & Malerba, D. (2020, September). Predictive process mining meets computer vision. In *International Conference on Business Process Management* (pp. 176-192). Springer International Publishing.

Petersen, K., Vakkalanka, S., & Kuzniarz, L. (2015). Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology*, *64*, 1-18.

Rehse, J. R. (2024). On process discovery experimentation: Addressing the challenges. *ACM Transactions on Software Engineering and Methodology*. https://doi.org/10.1145/3672447

Salehi, A., Aghdasi, M., Khatibi, T., & Sheikhmohammadi, M. (2023). Data quality in process mining: A systematic review. J. *Science & Tech. Info. Mgmt*.

Tax, N., & Van der Aalst, W. M. P. (2016). A machine learning approach to business process management. *Business & Information Systems Engineering, 58*(2), 101-110. https://doi.org/10.1007/s12599-016-0432-9

Van der Aalst, W. M. P. (2016). *Process mining: Data science in action* (2nd ed.). Springer. https://doi.org/10.1007/978-3-319-41535-5

Van der Aalst, W. M. P., Weijters, A. J. M. M., & Maruster, L. (2004). Workflow mining: Discovering process models from event logs. *IEEE Transactions on Knowledge and Data Engineering, 16*(9), 1128-1142. https://doi.org/10.1109/TKDE.2004.51

Van Der Aalst, W., Adriansyah, A., De Medeiros, A. K. A., Arcieri, F., Baier, T., Blickle, T., ... & Wynn, M. (2011, August). Process mining manifesto. In *International Conference on Business Process Management* (pp. 169-194). Springer Berlin Heidelberg.

van der Aalst, W. M. P. (2016). *Scalable process discovery and conformance checking*. Software and Systems Modeling.

Van Zelst, S., et al. (2018). On-the-fly process discovery: A new framework for process mining. In *10th International Conference on Business Process Management* (pp. 1-16). Springer. https://doi.org/10.1007/978-3-030-00450-3_1.

Vecino Sato, D.M., de Freitas, S.C., Barddal, J.P., & Scalabrin, E.E. (2021). A survey on concept drift in process mining. *arXiv*, 2112. 02000.

Verbeek, H. M. W., & de Carvalho, R. M. (2018). *Log skeletons: A classification approach to process discovery. arXiv*, *1806*. 08247.

Weijters, A., & Van der Aalst, W. M. P. (2003). A genetic algorithm for process mining. In *4th International Conference on Practical Applications of Knowledge Discovery and Data.*

Yang, Y., Wu, Z., Chu, Y., Chen, Z., Xu, Z., & Wen, Q. (2024). Intelligent cross-organizational process mining: A survey and new perspectives. *arXiv*, *2407*.11280.