



Data Quality in Process Mining: A Systematic Review

Ahmad Salehi^{ID}

Ph.D., Student, Department of Information Technology Engineering, Faculty of Industrial and Systems Engineering, Tarbiat Modares University, Tehran, Iran (**Corresponding author**).
ahmad.salehi@modares.ac.ir

Mohammad Aghdasi^{ID}

Professor, Department of Systems and Productivity Management, Faculty of Industrial and Systems Engineering, Tarbiat Modares University, Tehran, Iran. aghdasim@modares.ac.ir

Toktam Khatibi^{ID}

Associate Professor, Department of Industrial Engineering, Faculty of Industrial and Systems Engineering, Tarbiat Modares University, Tehran, Iran. toktam.khatibi@modares.ac.ir

Majid Sheikh Mohammadi^{ID}

Associate Professor, Department of Socio-economic Systems, Faculty of Industrial and Systems Engineering, Tarbiat Modares University, Tehran, Iran. msheikhm@modares.ac.ir

Abstract

Purpose: Process mining connects the disciplines of data mining and machine learning to business process management techniques. A business process is a series of independent and interdependent activities that transform inputs (data, materials, etc.) using one or more resources (such as time, employees, and money). It utilizes the necessary outputs. It is possible to examine the actual behavior of organizations, including the performance of individuals, departments, and resources, using process analysis techniques. The results of the process analysis, which typically includes the organization's business process models, can be compared to the organization's documents and requirements. Thus, processes will be able to be compared, reviewed, monitored, and enhanced. Process mining methods operate based on event logs stored in information systems. Using process mining without high-quality input data will not result in accurate conclusions about an organization's business processes. In recent years, researchers have focused on the evaluation and enhancement of the quality of input data using process mining techniques. The objective of this study is to identify and categorize the most significant data quality issues, as well as recognize the approaches proposed to address this challenge in process mining.

Methods: This research employs a systematic review with the intent of analyzing all valid evidence in order to answer the research questions. This study investigates 102 academic studies published between 2007 and 2021, including conference papers, journal articles, and theses. Towards this end, a systematic three-part research methodology was employed. In the first section, which included the research definition, the research field was defined first, followed by the research objectives and queries. In the concluding step of this section, the research's scope is defined. In the second section, the research methodology and entry criteria for the studies

Cite this article: Salehi, A., Aghdasi, M., Khatibi, T. & Sheikh Mohammadi, M. (2023). Data Quality in Process Mining: A Systematic Review. *Sciences and Techniques of Information Management*, 9(3): 103-160.
<https://doi.org/10.22091/STIM.2022.7800.1737>

Received: 2023-06-23 ; **Revised:** 2023-07-12 ; **Accepted:** 2023-07-22 ; **Published online:** 2023-07-25

© The Author(s).

Article type: Research

Published by: University of Qom.



discovered during the search for scientific resources are defined. Finally, the identified studies are evaluated in terms of their citations and classified. In the third section, which is devoted to the evaluation of the research, the concluding research of the study is conducted, and then, based on the investigation of the preceding studies, the findings and conclusions are determined. Important data and evidence were extracted from the collated research, allowing for the creation of the necessary tables and graphs.

Findings: In recent years, researchers have paid more attention to data quality challenges in the process mining, according to the findings of recent research. In 2019 and 2020, the greatest number of studies will have been published. It was also discovered that the majority of articles were published in three scientific databases, namely Springer, IEEE, and Elsevier. 51% of the studies examined were presented at prestigious conferences. 36% of the studies were published in prestigious scientific journals, while the remaining 13% were represented in dissertations and university reports. The study of the selected articles revealed that 20 data quality issues that can arise in the input data have been investigated in the literature. These challenges have been categorized into five levels: trace, event, case, activity, and timestamps, and four foundational approaches have been identified that have been used to evaluate and resolve data quality challenges in the mining process. 1) data quality frameworks 2) preprocessing 3) anomaly detection 4) repair. Our findings indicate that preprocessing techniques that seek to remove chaotic and infrequent behaviors from the event log have received more attention than other techniques. In addition, these results demonstrate that, in recent years, the discovery of anomalies and the reconstruction of missing events have become popular research topics within the field of process mining. Examining studies related to the field of data quality in the data mining process reveals an abundance of approaches and methods for addressing data quality challenges. Investigations revealed that the use of colorful Petri nets as a mathematical method has been considered in all selected research projects.

Conclusions: The data needed for process mining methods can be obtained from various sources. One of the major advantages of process mining is that it is not limited to a specific type of system. Any workflow-based system, such as ticketing, resource management, databases, data warehouses, legacy systems, and even manually collected data, can be analyzed as long as it can be separated using case ID, activity, and timestamp attributes. In real-world scenarios, most data is not collected for process mining purposes or is unsuitable for use in process mining analyses. Especially data that is recorded manually or scattered among various isolated systems can contain errors. Despite the efforts made to improve the quality of input data in the mining process, it is still necessary to develop efficient frameworks and methods to identify, evaluate, and address data quality challenges in real business processes, which are often characterized by high volume and complexity. The results of this research can offer a fresh perspective for researchers, data science specialists, and business analysts.

Keywords: Information Systems, Business Process Management, Process Mining, Data Quality, Event Log.



مرور نظام‌مند کیفیت داده در فرایند کاوی

احمد صالحی^{ID}

دانشجوی دکتری، گروه مهندسی فناوری اطلاعات، دانشکده مهندسی صنایع و سیستم‌ها، دانشگاه تربیت مدرس، تهران،
ahmad.salehi@modares.ac.ir (نویسنده مسئول)

محمد اقدسی^{ID}

استاد، گروه مدیریت سیستم و بهره‌وری، دانشکده مهندسی صنایع و سیستم‌ها، دانشگاه تربیت مدرس، تهران، ایران.
aghdasim@modares.ac.ir

توکتم خطیبی^{ID}

دانشیار، گروه مهندسی صنایع، دانشکده مهندسی صنایع و سیستم‌ها، دانشگاه تربیت مدرس، تهران، ایران.
toktam.khatibi@modares.ac.ir

مجید شیخ محمدی^{ID}

دانشیار، گروه مهندسی سیستم‌های اقتصادی و اجتماعی، دانشکده مهندسی صنایع و سیستم‌ها، دانشگاه تربیت مدرس، تهران، ایران.
msheikhm@modares.ac.ir

چکیده

هدف: فرایند کاوی ارتباط میان دو زمینه داده‌کاوی و یادگیری ماشین، با رویکردهای مدیریت فرایند کسب‌وکار را فراهم می‌سازد. یک فرایند کسب‌وکار ترتیبی از فعالیت‌های مستقل و به هم مرتبط است، که در هنگام انجام هر کدام از فعالیت‌های پیش‌بینی شده، از یک یا چند منبع (شامل زمان، افراد، هزینه و ...) برای تبدیل ورودی‌ها (داده، مواد و ...) به خروجی‌های مورد نیاز استفاده می‌کند. با استفاده از روش‌های فرایند کاوی می‌توان رفتار حقیقی سازمان‌ها شامل عملکرد افراد، بخش‌ها و منابع را بررسی نمود. نتایج حاصل از فرایند کاوی که به صورت معمول شامل مدل‌های فرایند کسب‌وکار است را می‌توان با مستندات و الزامات سازمان‌ها مقایسه کرد. بدین‌سان امکان بررسی، نظارت و بهبود فرایندها امکان‌پذیر خواهد شد. مبنای عمل روش‌های فرایند کاوی براساس نگاره‌های رویداد ذخیره شده در سامانه‌های اطلاعاتی است. استفاده از فرایند کاوی بدون وجود داده‌های ورودی باکیفیت، به یافته‌هایی باورپذیر در مورد فرایندهای کسب‌وکار یک سازمان منجر نخواهد شد. از این‌رو در سال‌های گذشته ارزیابی و افزایش کیفیت داده‌های ورودی به روش‌های فرایند کاوی، مورد توجه پژوهشگران قرار گرفته است. در این راستا، هدف پژوهش حاضر شناخت و گروه‌بندی مهم‌ترین مشکلات کیفیت داده و بازشناخت رویکردهای ارائه شده برای حل این چالش در فرایند کاوی است.

روش: روش مورد استفاده در این پژوهش از نوع مرور نظام‌مند بوده که با هدف واکاوی همه شواهد معتبر، برای پاسخگویی به سوالات پژوهش مورد استفاده قرار گرفته است. در این مطالعه به بررسی و کاوش ۱۰۲ مطالعه دانشگاهی در بین سال‌های ۲۰۰۷ تا ۲۰۲۱ میلادی پرداخته شده است که شامل پژوهش‌های انتشار یافته در همایش‌ها، مجلات و همچنین تعدادی پایان‌نامه می‌باشند. بدین منظور از یک روش پژوهش نظام‌مند سه بخشی استفاده شده است. در بخش اول

استناد به این مقاله: صالحی، ا.، اقدسی، م.، خطیبی، ت.، شیخ محمدی، م. (۱۴۰۲). مرور نظام‌مند کیفیت داده در فرایند کاوی. *علوم و فنون مدیریت*

اطلاعات، ۹(۳): ۱۶۰-۱۰۳. <https://doi.org/10.22091/STIM.2022.7800.1737>

تاریخ دریافت: ۱۴۰۲/۰۴/۰۲ ؛ تاریخ اصلاح: ۱۴۰۲/۰۴/۲۱ ؛ تاریخ پذیرش: ۱۴۰۲/۰۴/۳۱ ؛ تاریخ انتشار آنلاین: ۱۴۰۲/۰۵/۰۳

ناشر: دانشگاه قم

نوع مقاله: پژوهشی

© نویسندگان.



که شامل تعریف پژوهش بوده، ابتدا زمینه پژوهش تعریف شده و سپس اهداف و پرسش‌های پژوهش مشخص شده‌اند. در گام نهایی این بخش، دامنه پژوهش تعیین گردیده است. در بخش دوم، روش‌شناسی پژوهش معین شده و معیارهایی برای وارد کردن مطالعات یافت شده به فرایند بررسی نهایی، مشخص شده‌اند. سپس این مطالعات از لحاظ ارجاعات مورد بررسی قرار گرفته و در پایان دسته‌بندی شده‌اند. در بخش سوم که به ارزیابی پژوهش‌های جمع‌آوری شده اختصاص یافته است، پژوهش‌های نهایی مطالعه شده و سپس بر مبنای کاوش آن‌ها، یافته‌ها و نتایج حاصل تعیین گردیده‌اند. با بررسی جامع پژوهش‌های گردآوری شده، داده‌ها و شواهد مهم استخراج گردیده و جداول و نمودارهای مورد نیاز نیز رسم شده‌اند.

یافته‌ها: براساس بررسی پژوهش‌های انجام شده، توجه به چالش‌های کیفیت داده در فرایندکاوی در سال‌های اخیر افزایش یافته است. بیشترین حجم از مطالعات در سال‌های ۲۰۱۹ و ۲۰۲۰ میلادی منتشر شده‌اند. همچنین مشخص شد که بیشتر مقالات به ترتیب در سه پایگاه علمی Elsevier، Springer، و IEEE انتشار یافته‌اند. ۵۱ درصد از مطالعات مورد بررسی در همایش‌ها و کنفرانس‌های معتبر ارائه شده‌اند. ۳۶ درصد از پژوهش‌ها در مجلات معتبر علمی انتشار یافته و ۱۳ درصد از مطالعات از طریق پایان‌نامه‌ها و گزارشات دانشگاهی انعکاس یافته‌اند. برآوردهای حاصل از مطالعه مقالات منتخب نشان داد که ۲۰ مشکل کیفیت داده که می‌توانند در داده‌های ورودی پدید آیند، در ادبیات مورد کنکاش قرار گرفته‌اند. این چالش‌ها در پنج سطح مسیر فرایند، رویداد، مورد، فعالیت و برچسب زمانی طبقه‌بندی شده و چهار رویکرد بنیادین که برای ارزیابی و حل چالش‌های کیفیت داده در فرایندکاوی به کار گرفته شده‌اند، مشخص گردیدند که شامل: چارچوب‌های ارزیابی کیفیت داده، پیش‌پردازش، کشف ناهنجاری، و ترمیم می‌باشند. یافته‌ها نشان می‌دهد که روش‌های پیش‌پردازش که هدف آن‌ها حذف رفتارهای آشفته و کم تکرار از داده‌های ورودی است، بیشتر از دیگر رویکردها مورد توجه واقع شده‌اند. همچنین در سال‌های اخیر کشف ناهنجاری و بازسازی رویدادهای گمشده به جستاری جذاب در زمینه فرایندکاوی تبدیل شده‌اند. یکی دیگر از نتایج مهم، بررسی مطالعات مربوط به زمینه کیفیت داده در فرایندکاوی، استفاده از حجم وسیعی از رویکردها و روش‌ها برای حل چالش‌های کیفیت داده است. بررسی‌ها نشان داد که استفاده از شبکه‌های پتری، نقش پر رنگی در همه پژوهش‌های منتخب به عنوان یک رهیافت ریاضی مد نظر پژوهشگران داشته است.

نتیجه‌گیری: داده‌های مورد نیاز روش‌های فرایندکاوی می‌تواند از مکان‌های مختلفی به دست آید. یکی از مزایای بزرگ فرایندکاوی این است که مختص به یک نوع سامانه خاص نیست. هر سامانه مبتنی بر گردش کار، فروش بلیط، مدیریت منابع، پایگاه و انبارهای داده، سامانه‌های قدیمی و حتی داده‌هایی که به صورت دستی جمع‌آوری شده‌اند، تا زمانی که با استفاده از ویژگی‌های شناسه مورد، فعالیت و برچسب زمانی قابل تفکیک باشند، قابل تجزیه و تحلیل هستند. اما در مسائل دنیای واقعی، بیشتر داده‌ها برای اهداف فرایندکاوی جمع‌آوری نشده‌اند یا مناسب استفاده برای تحلیل‌های فرایندکاوی نیستند. به خصوص داده‌هایی که به صورت دستی ثبت شده‌اند یا داده‌هایی که در بین سامانه‌های جزیره‌ای مختلف پراکنده شده‌اند، می‌توانند حاوی خطا باشند. با وجود تلاش‌های مناسب برای بهبود کیفیت داده‌های ورودی در فرایندکاوی، همچنان توسعه چارچوب‌ها و روش‌های کارآمد با هدف شناسایی، ارزیابی و حل چالش‌های کیفیت داده در فرایندهای کسب‌وکار واقعی که بیشتر اوقات از حجم و پیچیدگی بالایی برخوردارند، پیشنهاد می‌گردد. نتایج این پژوهش می‌تواند به عنوان دورنمایی جدید برای پژوهشگران و متخصصان علم داده و تحلیل‌گران کسب‌وکار مورد استفاده قرار گیرد.

کلیدواژه‌ها: سامانه‌های اطلاعاتی، مدیریت فرایندهای کسب‌وکار، فرایندکاوی، کیفیت داده، گزارش رویداد.

۱. مقدمه

پیدایش مفهوم فناوری اطلاعات و گسترش ابزارهای وابسته به آن منجر به ایجاد دگرگونی‌های بنیادین در نحوه طراحی، پیاده‌سازی و پشتیبانی از فرایندهای کسب‌وکار در سازمان‌ها و شرکت‌های تجاری شده است. امروزه سازمان‌ها دریافته‌اند که فناوری اطلاعات دیگر یک پشتیبان یا توانمندساز ساده نخواهد بود، بلکه فناوری اطلاعات و به دنبال آن فناوری دیجیتال خود در مرکز کسب‌وکارهای امروزی قرار دارند و با وجود آن‌ها است که امکان ارزش‌آفرینی فراهم خواهد شد (شامی زنجانی، نبی و ایران‌دوست، ۱۳۹۹). از این‌رو با حرکت هرچه بیشتر سازمان‌ها به سمت دیجیتالی شدن، اهمیت همپوشانی و درهم تنیدگی سامانه‌های اطلاعاتی با فرایندهای کسب‌وکار آشکارتر گردیده (Li, 2020; Van Der Aalst, 2016) و موجب شده است که فرایندها و داده‌های ذخیره شده درون بانک‌های اطلاعاتی، به دارایی‌هایی ارزشمند برای سازمان‌ها و کسب‌وکارها تبدیل شوند (Dumas & et al., 2018; Laranjeiro & et al., 2015).

فرایندکاوی^۱ به عنوان یک رویکرد داده‌محور^۲، به منظور استخراج دانش نهان از داده‌های پدید آمده به هنگام اجرای فرایندها، معرفی و توسعه یافته است. فرایندکاوی با استفاده از ترکیب روش‌های «یادگیری ماشین»^۳ و مفاهیم «مدیریت فرایندهای کسب‌وکار»^۴ به دنبال کشف و بهبود فرایندها از طریق کاوش داده‌های ذخیره شده در پایگاه داده سامانه‌های اطلاعاتی است (Zerbino & et al., 2021). روش‌های فرایندکاوی سه هدف اصلی را پیگیری می‌کنند:

(۱) کشف مدل‌های فرایند^۵ کسب‌وکار از داده‌های ورودی،

(۲) بررسی همانندی^۶ مدل کشف شده با قوانین کسب‌وکار و مستندات سازمان،

(۳) ارائه رهیافت‌هایی برای بهبود^۷ فرایندهای کنونی (Erdogan & Tarhan, 2018). برای استفاده از مزیت‌های فرایندکاوی باید در ابتدا داده‌های مورد نیاز را براساس اهداف در نظر گرفته شده، از بانک‌های اطلاعاتی واکنشی^۸ نمود. اما با توجه به بلوغ سازمانی و همچنین وضعیت حاکمیت داده و

1. Process mining
2. Data driven
3. Machine learning
4. Business process management (BPM)
5. Process discovery
6. Conformance checking
7. Enhancement
8. Fetch

فناوری اطلاعات در سازمان، سطوح ذخیره‌سازی اطلاعات و رفتار ناشی از اجرای فرایندها ممکن است متفاوت باشد. در سازمان‌هایی با بلوغ بالا، همه روندهای اجرایی به صورت الکترونیکی و با مدیریت سامانه‌های اطلاعاتی انجام می‌شوند، اما در سازمان‌هایی با بلوغ کمتر، امکان دارد اجرای فعالیت‌ها به صورت ترکیبی از فرایندهای الکترونیکی و دستی باشد (Van Der Aalst & et al., 2011). در این حالت ممکن است بخشی از اطلاعات در دفاتر مخصوص ذخیره شده و چه بسا هرگز ثبت نگردند. بنابراین، پیگیری آنچه که به راستی در سازمان رخ داده است، آسان نخواهد بود.

بدین‌سان بررسی و ارزیابی کیفیت داده‌های در دسترس برای سازمان‌ها از اهمیت زیادی برخوردار خواهد بود. در مطالعه Commission (۲۰۰۸) کیفیت داده اینگونه تعریف شده است: «به میزان یا درجه‌ای که ویژگی‌های داده می‌توانند الزامات مورد نیاز را برآورده نمایند». داده‌های با کیفیت پایین می‌تواند بر روی عملکرد کارکنان، تصمیمات و استراتژی‌های یک سازمان اثر منفی بگذارد و منجر به از دست رفتن فرصت‌های رقابتی گردد (Redman, 1998)؛ زیرا داده‌های جمع‌آوری شده از دنیای واقعی اغلب ناخالص^۱، ناسازگار^۲، غیردقیق^۳، ناکامل^۴، قدیمی^۵ یا تکراری^۶ هستند و وجود چنین داده‌هایی برای یک سازمان یا کسب‌وکار هزینه‌بر خواهد بود (Fan, 2012). برابر با گزارش‌های منتشر شده در صنایع، سرمایه‌گذاری زیادی برای بهبود کیفیت داده‌های مورد استفاده صورت می‌پذیرد، زیرا با توجه به تنوع ابعاد، ویژگی‌های داده‌ها، دیدگاه‌های کسب‌وکار و همچنین مشخصات سامانه‌هایی که مدیریت داده‌ها را برعهده دارند، درک ابعاد و اندازه‌گیری کیفیت داده‌ها می‌تواند یک فعالیت بسیار دشوار باشد (Laranjeiro & et al., 2015). در نتیجه واکاوی‌های انجام گرفته بر پایه داده‌های ورودی دارای مشکل، غیرقابل اعتماد خواهند بود. به بیانی دیگر، داده‌های ورودی نمی‌توانند بازنمایی درستی از واقعیت رخ داده ارائه دهند

(Bose & et al., 2013; Cheng & Kumar, 2015; Kherbouche & et al., 2016; Martin, 2021; Sani & et al., 2018a).

به موازات جذاب شدن استفاده از روش‌های فرایندکاوی در پژوهش‌های علمی و صنعتی

1. Dirty
2. Inconsistent
3. Inaccurate
4. Incomplete
5. Stale
6. Duplicate

در سطح جهان (Reinkemeyer, 2020; Zerbino & et al., 2021)، پژوهشگران ایرانی نیز از روش‌های فرایندکاوی در زمینه‌های مختلف از جمله صنعت بانکداری (مصطفایی دولت و همکاران، ۱۳۹۸؛ خوش‌خوی نیلاش، تمجیدیا مچلو و راد، ۱۴۰۰)، صنعت بیمه (خدیور، فیروزی و نیاکان، ۱۳۹۹)، فرایندهای دانشگاهی (حسینی، مصلح و حسینی، ۱۳۹۷)، مدیریت منابع سازمانی (علیزاده و هادوی‌نژاد، ۱۳۹۸) و همچنین توسعه روش‌های نوین فرایندکاوی (خجسته، کاهانی و بهکمال، ۱۳۹۹) استفاده کرده‌اند. از این‌رو با توجه به نبود مطالعات مناسب برای بررسی چالش‌های کیفیت پایین داده در فرایندکاوی به زبان فارسی، ضرورت انجام یک پژوهش جامع برای بررسی و معرفی موارد مربوط به این چالش را بایسته می‌سازد. هدف پژوهش حاضر مرور نظام‌مند مطالعات انجام شده در زمینه فرایندکاوی، با تمرکز بر ارزیابی و بهبود کیفیت داده‌ها می‌باشد.

۲. پیش‌نیازها

ورودی روش‌های فرایندکاوی، نوع ویژه‌ای از داده‌ها به نام «گزارش رویداد»^۱ یا به اختصار «لاگ»^۲ است^۳ (ون در آلتست، ۱۳۹۴). گزارش‌های رویداد، منابع بسیار باارزشی در مورد اجرای فرایندهای کسب‌وکار هستند (Bayomie & et al., 2016). گزارش‌های رویداد می‌توانند به صورت مستقیم توسط سامانه‌های اطلاعاتی فرایند- آگاه^۴ تولید شده (Dumas & et al., 2018)، یا از طریق ترکیب چند منبع داده‌ای توسط فرایندهایی موسوم به استخراج، تبدیل و بارگذاری^۵ ایجاد شوند (Van Der Aalst & et al., 2011).

جدول شماره (۱) نمونه‌ای از یک گزارش رویداد استاندارد که می‌تواند به عنوان ورودی، به تحلیل‌های فرایندکاوی وارد شود (به عنوان نمونه کشف فرایند) را نشان می‌دهد. همان‌طور که مشاهده می‌شود، جدول مورد اشاره دارای تعدادی ردیف^۶ و ستون^۷ (ویژگی) است. یک گزارش رویداد استاندارد باید حداقل دارای سه ستون (ویژگی) زیر باشد، در غیر این صورت امکان استفاده از روش‌های فرایندکاوی بر روی آن امکان‌پذیر نخواهد بود (Van Der Aalst, 2016). این سه ویژگی

1. Event log

2. Log

۳. در برخی از متون فارسی این عبارت تحت عنوان «نگاره رویداد» ترجمه شده است.

4. Process-aware information systems (PAIS)

5. Extract, transform, load (ETL)

6. Record

7. Column or attribute

عبارتند از:

- **مورد یا نمونه فرایند:**^۱ برای مشخص کردن اینکه هر فعالیت برای کدام مورد (نمونه) در حال انجام است (در جدول (۱) دو مورد یا نمونه فرایند (۱ و ۲) وجود دارد).
- **وظیفه^۲ یا فعالیت:**^۳ این ویژگی، نام کار انجام شده در هر مرحله از فرایند را مشخص می‌نماید.
- **برچسب زمانی:**^۴ این ویژگی، تاریخ و زمان انجام هر فعالیت را تعیین می‌نماید.

جدول ۱- نمونه‌ای از یک گزارش رویداد

Event_Id	Case_Id	Task	Timestamp
1001	1	a	8/30/2021 08:43
1002	1	b	8/30/2021 09:21
1003	1	c	8/30/2021 08:40
1004	1	e	8/30/2021 10:11
1005	2	a	9/03/2021 11:23
1006	2	b	9/03/2021 13:01
1007	2	d	9/03/2021 09:30
1008	2	e	9/03/2021 14:20

در این جدول، ستونی به نام شناسه رخداد^۵ هم وجود دارد که برای ایجاد تمایز بین رویدادهای ذخیره شده در یک لاگ استفاده می‌شود.

گزارش رویداد علاوه بر شناسه رخداد می‌تواند ویژگی‌های اختیاری دیگری مانند هزینه، بخش و... هم داشته باشد (Van Der Aalst, 2016) که برای دیگر تحلیل‌های فرایندکاوی مناسب هستند (به عنوان نمونه، تحلیل هزینه منابع یا تحلیل شبکه اجتماعی موجود در یک سازمان برای یک فرایند کسب‌وکار خاص). به مقادیر ذخیره شده برای هر ویژگی، برچسب^۶ گفته می‌شود. برای اینکه الگوریتم‌های متداول کشف فرایند مانند الگوریتم آلفا (Van der Aalst & et al., 2004) بتوانند عملیات کشف فرایند را به صورت صحیح انجام دهند، نیاز است که ترتیب انجام فعالیت‌ها به

1. Case_id
2. Task
3. Activity
4. Timestamps
5. Event_Id
6. Label

صورت صحیح در گزارش رویداد، ذخیره شده باشد (Van Der Aalst, 2011) که این مورد با استفاده از ویژگی «برچسب زمانی» کنترل می‌گردد.

در ادامه این بخش، چکیده مطالب مهم در مورد یک گزارش رویداد آمده که برای دانستن مشکلات کیفیت داده در فرایندکاوی بسیار با اهمیت است (Bose, 2012):

- گزارش رویداد شامل اطلاعات ثبت شده ناشی از اجرای یک فرایند کسب‌وکار است.
- هر گزارش رویداد شامل تعدادی نمونه فرایند است. در جدول (۱)، دو مورد وجود دارند که با شناسه (۱ و ۲) مشخص شده‌اند.

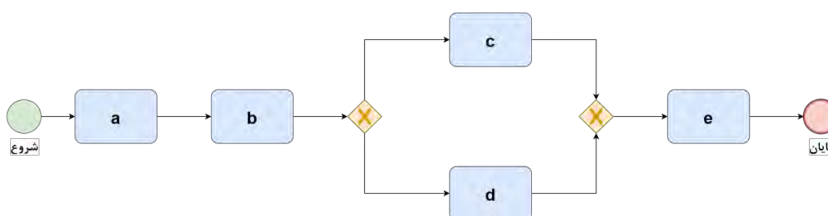
● برای هر مورد، ترتیب خاصی از رویدادها انجام می‌گیرد. اگر فرض کنیم که هر رویداد، معادل با یک فعالیت باشد، رویدادهای یک نمونه فرایند، به دنباله‌ای از فعالیت‌ها تبدیل خواهد شد که به آن‌ها یک «مسیر»^۱، «مسیر فرایند» یا «دنباله»^۲ می‌گویند. به عنوان نمونه مسیر مورد (۱) شامل فعالیت‌های $\{a, b, c, e\}$ و مسیر مورد (۲) شامل فعالیت‌های $\{a, b, d, e\}$ می‌شوند که به وسیله ویژگی برچسب زمانی مرتب شده‌اند.

- هر رویداد متعلق به یک نمونه فرایند است.

- رویدادها ویژگی‌های مختلفی دارند (فعالیت، برچسب زمانی و...).

نمودار (۱) مدل فرایند کسب‌وکار کشف شده بر مبنای رویدادهای ذخیره شده در جدول شماره (۱) را نشان می‌دهد. در ادبیات فرایندکاوی از دو شیوه متداول «شبکه‌های پتری»^۳ یا «مدل و نشانه‌گذاری فرایند کسب و کار»^۴ (BPMN) برای ترسیم مدل‌های فرایند استفاده می‌شود (Burattin, 2015). در این پژوهش از استاندارد BPMN 2.0 استفاده شده است. در این نمودار فعالیت‌ها با مستطیل‌های آبی رنگ مشخص شده‌اند. همچنین نقاط تصمیم‌گیری با استفاده از لوزی‌های زرد رنگ به تصویر کشیده شده‌اند که نشان‌دهنده انجام فعالیت‌های متمایز برای نمونه فرایندهای متفاوت براساس قوانین کسب‌وکار است. شروع و پایان فرایند هم به ترتیب با استفاده از دایره‌های سبز و قرمز رنگ تعیین شده‌اند. نمودار (۱) نمونه‌ای از فرایند کشف شده از یک گزارش رویداد براساس روش‌های کشف فرایند در فرایندکاوی است.

1. Trace
2. Sequence
3. Petri nets
4. Business Process Model and Notation



نمودار ۱- مدل فرایند استخراج شده از رویدادهای ذخیره شده در جدول (۱)

۳. روش پژوهش

برای انجام این پژوهش از الگوی ارائه شده در مطالعات (Al-Hashedi & Magalingam, 2021; Ngai & et al., 2011) استفاده شده است. بدین ترتیب مدل مفهومی پژوهش حاضر مشتمل بر سه بخش اصلی است، شامل تعریف پژوهش، روش‌شناسی پژوهش و ارزیابی پژوهش. فرایند انجام این پژوهش در نمودار (۲) ترسیم شده است.

در بخش تعریف پژوهش، سه فعالیت کلیدی انجام شده است. اولین گام، تعیین زمینه پژوهش بوده که شامل بررسی مفهوم کیفیت داده در فرایندکاوی است. فعالیت دوم، مشخص نمودن اهداف و پرسش‌های پژوهش زمینه مورد مطالعه است. هدف این پژوهش، بررسی و دسته‌بندی مقالات ارائه شده مبتنی بر زمینه تحقیق و شناخت آخرین روش‌های معرفی شده است. همچنین سوالات مورد بررسی در پژوهش حاضر در جدول (۲) آمده است.

جدول ۲- پرسش‌های پژوهش

ردیف	پرسش پژوهش
۱	نحوه توزیع پژوهش‌های ارائه شده برای بهبود کیفیت داده‌ها در فرایندکاوی، در پایگاه‌های علمی، نشریات و همایش‌ها به چه صورت است؟
۲	انواع مشکلات کیفیت داده که در ادبیات مورد توجه قرار گرفته‌اند، شامل چه مواردی هستند؟
۳	انواع رهیافت‌های ارائه شده برای مقابله با چالش کیفیت داده در فرایندکاوی، بر چه رویکردهایی استوار می‌باشند؟
۴	انواع روش‌های مورد استفاده در پژوهش‌های مورد بررسی کدامند؟
۵	مطالعات بررسی شده برای پیاده‌سازی رویکرد خود، از کدام چارچوب‌ها و ابزارهای نرم‌افزاری بهره برده‌اند؟
۶	پژوهش‌های انتشار یافته برای ارزیابی روش‌های خود از کدام مجموعه داده‌های استاندارد استفاده نموده‌اند؟

فعالیت سوم، تعریف دامنه پژوهش بوده که شامل بررسی پژوهش‌هایی است که در بازه زمانی سال‌های ۲۰۲۱-۲۰۰۷ به رشته تحریر درآمده و به بررسی و بهبود کیفیت داده در فرایندکاوی پرداخته‌اند.

در بخش دوم که شامل روش‌شناسی پژوهش می‌باشد، چهار فعالیت و سه نقطه تصمیم‌گیری تعیین شده است. اولین فعالیت، جستجوی پایگاه‌های علمی برای یافتن مقالات مربوط به زمینه و هدف تحقیق می‌باشد.

از این‌رو منابع علمی شناخته شده براساس دسترسی نویسندگان به منابع علمی، مورد جستجو قرار گرفته‌اند که شامل پایگاه‌های علمی Google Scholar، ACM، IEEE، Springer، Science Direct می‌باشند. مبنای انتخاب بانک‌های اطلاعاتی اشاره شده براساس بررسی و شناخت مهم‌ترین تارنماهای علمی و نشریاتی است که نتایج حاصل از پژوهش‌های بدیع مرتبط با زمینه فرایندکاوی، تاکنون در آن‌ها منتشر شده است. برای دستیابی به این هدف، در واری‌های ابتدایی، فهرستی از مهم‌ترین کنفرانس‌ها و نشریاتی که اقدام به انتشار مقالات فرایندکاوی نموده‌اند، آماده گردید که بیشترین ارتباط را با پایگاه‌های علمی مورد اشاره داشته‌اند. از طرفی، براساس «بیانیه فرایندکاوی»^۱ که توسط «کارگروه ویژه فرایندکاوی در موسسه مهندسان برق و الکترونیک» انتشار یافته است،^۲ فهرستی از مهم‌ترین پژوهشگران فعال در زمینه فرایندکاوی مشخص شده است که با استفاده از داده‌های تارنمای Google Scholar، پایگاه‌های علمی مهمی که نتایج حاصل از مطالعات این پژوهشگران در آن‌ها منتشر شده است، شناسایی گردیدند. برای کسب اطمینان بیشتر، با استفاده از تحلیل ارجاعات،^۳ دیگر نشریات مهم که توسط برخی دیگر از بانک‌های اطلاعاتی از جمله Web Of Science و همچنین Scopus نمایه شده‌اند، نیز شناسایی شدند.

برای یافتن پژوهش‌های موردنظر، از واژگان کلیدی مانند «کیفیت داده و فرایندکاوی»، «کیفیت گزارش رویداد»، «پیش‌پردازش در فرایندکاوی»، «چارچوب‌های کیفیت داده در فرایندکاوی»، «ارزیابی کیفیت داده در فرایندکاوی»، «کشف ناهنجاری در گزارش رویداد» و «بازسازی و ترمیم گزارش رویداد» استفاده شده است. همچنین با استفاده از عملگرهای بولی،^۴ برای رسیدن به نتایج موردنظر، کلیدواژه‌های تعیین شده با یکدیگر ترکیب و رشته‌های^۵ ایجاد شده، در پایگاه‌های منتخب مورد جستجو قرار گرفتند. در جدول (۳) فهرست کاملی از واژگان، واژگان مترادف یا جایگزین، به همراه رشته‌های جستجو گزارش شده‌اند.

<http://stlm.gom.ac.ir>

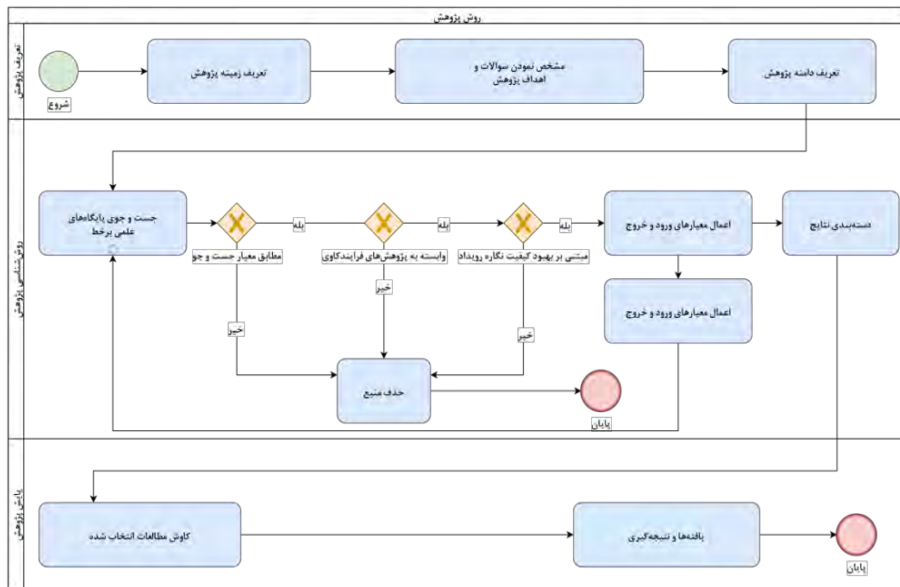
1. Process Mining manifesto
2. IEEE Task Force on Process Mining (<https://www.tf-pm.org>)
3. Citation analysis
4. Boolean Operator
5. String

جدول ۳ - کلیدواژه‌ها و رشته‌های جستجو شده

جایگزین‌ها	واژه مترادف	واژه یا اصطلاح جستجو شده
		Process Mining
Data Quality Assessment and Process Mining		Data Quality and Process Mining
	Log Quality	Event Log Quality
		Data Quality Framework in Process Mining
Pre-processing		Preprocessing
		Filtering
		Refinement
Event log Simplification		Event Log Abstraction
		Event Matching
		Activity Matching
Imputation		Repairing
Sequence		Trace
Infrequent behavior	Noisy	Noise
		Imperfection
		Anomaly Detection
Event log Simplification		Event Log Abstraction
رشته‌های جستجو		
("Data Quality" OR "Log Quality" OR "Data Quality Framework" OR "Data Quality Assessment") "Event Log" "Process Mining"		
("Filtering" OR "Cleaning" OR "Repairing" OR "Refinement" OR "Preprocessing" OR "Pre-processing") "Event Log" "Process Mining"		
("filtering" OR "cleaning" OR "repairing" OR "Refinement" OR "Preprocessing" OR "Pre-processing") "Trace" "Sequence" "Process Mining"		
("Anomaly detection" OR "Noise" "Infrequent behavior" OR "Noisy" OR "Imperfection") "Event log" "Process Mining"		
("Anomaly detection" OR "Noise" "Infrequent behavior" OR "Noisy" OR "Imperfection") "Trace" "Process Mining"		

همان‌طور که در نمودار (۲) مشخص شده است، سه نقطه تصمیم‌گیری برای هر پژوهش‌کاو شده وجود دارد. دروازه تصمیم‌گیری اول به این نکته می‌پردازد که آیا مطالعه یافت شده، یک پژوهش علمی است یا خیر؟ بعضی از اسناد منتشر شده، مربوط به حوزه‌های تحقیقات عملی نبوده که از فرایند تحقیق حذف شدند. دومین دروازه، مرتبط بودن مطالعه به زمینه فرایندکاوی را بررسی می‌نماید. بعضی از پژوهش‌ها به کیفیت داده در لاگ از دیدگاه‌های دیگر که مربوط به زمینه فرایندکاوی نیست، پرداخته‌اند که از بررسی حذف شدند. دروازه مهم بعدی این است که مقاله به یکی از جنبه‌های

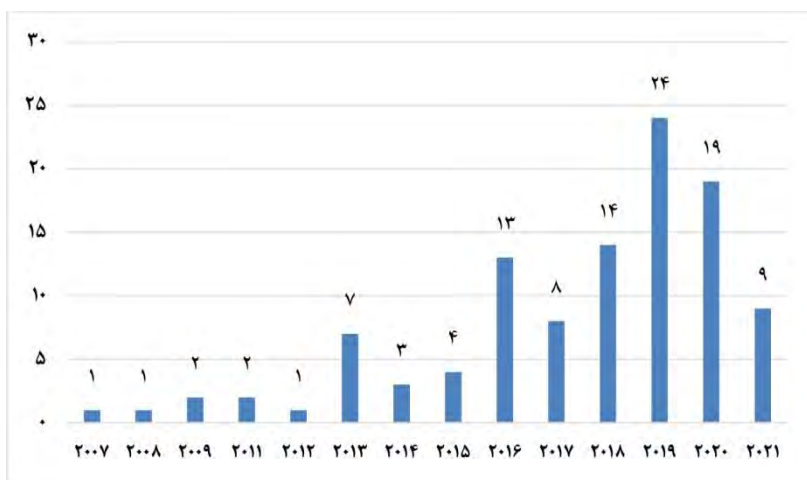
کیفیت داده در فرایندکاوی اشاره داشته باشد. بعضی از مطالعات مربوط به کیفیت مدل فرایند و بررسی کیفیت گزارش رویداد در زمینه «کشف فرایند» یا «بررسی همانندی» بوده است که مرتبط با سوالات پژوهش این مطالعه نبوده‌اند. فقط پژوهش‌هایی از زمینه کشف فرایند در این تحقیق در نظر گرفته شده‌اند که به کیفیت داده به عنوان یک روش پیش‌پردازش نگاه کرده باشند. همچنین مقالات مربوط به روش‌های کشف ناهنجاری، به گونه‌ای انتخاب شده‌اند که مربوط به زمینه ادبیات کیفیت داده باشند؛ زیرا بعضی از مطالعات به جنبه‌های دیگر فرایندکاوی اشاره داشته‌اند که خارج از دامنه این پژوهش می‌باشند. پس از این سه نقطه تصمیم‌گیری، فعالیت ورود و خروج مقالات به این مطالعه، براساس معیارهای تعیین شده است. در این مرحله مطالعات تکراری از فرایند بررسی حذف شده‌اند. همچنین تلاش شده است که کیفیت همایش یا مجله‌ای که مقاله در آن منتشر شده، مورد توجه قرار گیرد. بدین گونه ۱۰۲ پژوهش وارد فرایند مطالعه و ارزیابی شدند که بیشتر پژوهش‌ها در سال‌های ۲۰۱۹ و ۲۰۲۰ میلادی انتشار یافته‌اند. نمودار (۳) تعداد پژوهش‌های مورد بررسی را به تفکیک سال انتشار نشان می‌دهد. همچنین باید گفت که در گام جستجو، هیچ‌گونه پژوهش نگارش‌یافته به زبان فارسی، مطابق با اهداف این پژوهش یافت نشد.



نمودار ۲- فرایند انجام پژوهش

در بخش سوم که مربوط به مرحله ارزیابی پژوهش است، پژوهش‌های مورد مطالعه وارد گام

ارزیابی شده و بر مبنای سوالات پژوهش و استخراج داده‌ها، به دسته‌بندی‌های کاوش شده تقسیم شده‌اند. در فعالیت نهایی تلاش شده است براساس نتایج به دست آمده از بررسی و دسته‌بندی پژوهش‌های منتخب، روندهای زمینه مورد مطالعه کشف و پیشنهادهایی برای فعالیت‌های آینده ارائه گردد.



نمودار ۳- تعداد پژوهش‌های مورد بررسی به تفکیک سال انتشار

۴. یافته‌ها

۴-۱. شناسایی پایگاه‌های علمی میزبان پژوهش‌های کیفیت داده در فرایند کاوی

برای پاسخ به پرسش شماره (۱)، بررسی‌ها نشان داد که پژوهش‌های صورت گرفته در پیوند با زمینه کیفیت داده در فرایند کاوی، در ۱۷ پایگاه علمی شامل نشریات، مجلات و تارنماهای دانشگاهی منتشر شده‌اند. مطابق انتظار، بیشتر مقالات به ترتیب در سه پایگاه علمی Springer، Elsevier و IEEE انتشار یافته‌اند. مقالات چاپ شده در سه پایگاه فوق، ۶۶ درصد از کل مطالعات انجام شده را دربرگرفته‌اند. منابع دانشگاهی شامل پایان‌نامه‌ها، مقالات یا گزارشات علمی، در رتبه چهارم قرار دارند. گفتنی است که از مجموعه ۱۰۲ پژوهش بررسی شده، ۵۲ مورد در همایش‌ها ارائه شده و ۳۷ مورد نیز در مجلات منتشر شده‌اند که به ترتیب ۵۱٪ و ۳۶٪ از کل پژوهش‌های مورد بررسی را شامل می‌شوند. همچنین ۵ پایان‌نامه و ۸ گزارش دانشگاهی در این زمینه انجام شده است که سهم آن‌ها برابر با ۱۳٪ مطالعات کاوش شده بوده و در نمودار (۴) مشخص گردیده‌اند. جدول (۴) نیز توزیع ۸۹ پژوهش منتشر شده در نشریات و همایش‌های مرتبط را

براساس عنوان نشریه یا همایش، فراوانی (تعداد پژوهش‌ها) و درصد (درصد تعداد پژوهش‌های کلی مربوط به یک نشریه یا همایش) نشان می‌دهد. بدین‌سان در بین همایش‌ها، کنفرانس "International Conference on Business Process Management" با ۱۵ پژوهش و در بین نشریات، مجله "IEEE Transactions on Knowledge and Data Engineering" با ۵ پژوهش، میزبان بیشترین مقالات در زمینه کیفیت داده در فرایندکاوی بوده‌اند.



نمودار ۴- انواع پژوهش‌های مورد مطالعه براساس نوع انتشار

جدول ۴- فهرست نشریات/ همایش‌های میزبان پژوهش‌های مورد بررسی

عنوان نشریه / همایش	تعداد	درصد (%)
International Conference on Business Process Management	۱۵	۱۶,۸۵
OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"	۶	۶,۷۴
International Conference on Process Mining (ICPM)	۵	۵,۶۱
IEEE Transactions on Knowledge and Data Engineering	۵	۵,۶۱
Information systems	۵	۵,۶۱
International Conference on Advanced Information Systems Engineering	۵	۵,۶۱
Enterprise, Business-Process and Information Systems Modeling	۳	۳,۳۷
Decision Support Systems	۲	۲,۲۴
Expert Systems with Applications	۲	۲,۲۴
IEEE International Conference on Data Engineering	۲	۲,۲۴
IEEE International Conference on Web Services (ICWS)	۲	۲,۲۴
International Conference on Service-Oriented Computing	۲	۲,۲۴
International journal of environmental research and public health	۱	۱,۱۲
ACM SIGAPP Applied Computing Review	۱	۱,۱۲
ACM SIGMOD International Conference on Management of Data	۱	۱,۱۲
BPM reports	۱	۱,۱۲

درصد (%) تعداد	عنوان نشریه / همایش
۱،۱۲	Business-Process and Information Systems Modeling
۱،۱۲	Connection Science
۱،۱۲	Data
۱،۱۲	Data & Knowledge Engineering
۱،۱۲	Enterprise Modelling and Information Systems Architectures (EMISAJ)
۱،۱۲	European Conference on Information Systems (ECIS2020)
۱،۱۲	Health informatics journal
۱،۱۲	IEEE Access
۱،۱۲	IEEE International Conference on Healthcare Informatics (ICHI)
۱،۱۲	IEEE Symposium Series on Computational Intelligence (SSCI)
۱،۱۲	IFIP Working Conference on The Practice of Enterprise Modeling
۱،۱۲	Information and Organization
۱،۱۲	Information Sciences
۱،۱۲	International Conference on Applications and Theory of Petri Nets and Concurrency
۱،۱۲	International Conference on Business Information Systems
۱،۱۲	International Conference on Computational Collective Intelligence
۱،۱۲	International Conference on Computational Intelligence and Security (CIS)
۱،۱۲	International Conference on Computer Science, Engineering and Applications (ICCSEA)
۱،۱۲	International Conference on Design Science Research in Information Systems
۱،۱۲	International Conference on Discovery Science
۱،۱۲	International Conference on Enterprise Information Systems
۱،۱۲	International Journal of Advanced Science and Technology
۱،۱۲	International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)
۱،۱۲	Journal of Ambient Intelligence and Smart Environments
۱،۱۲	Journal of Computers
۱،۱۲	Journal of Intelligent Information Systems
۱،۱۲	Knowledge and Information Systems
۱،۱۲	Machine Learning
۱،۱۲	Proceedings of the International Symposium on Automation and Robotics in Construction (ISARC)
۱،۱۲	Scientific Programming
۱،۱۲	Simulation Modelling Practice and Theory
۱۰۰	مجموع

۴-۲. انواع مشکلات کیفیت داده در گزارش رویداد

برای یافتن پاسخ پرسش شماره (۲)، نخست پژوهش‌های منتخب مطالعه و بررسی شده و مشکلات کیفیت داده مورد اشاره از آن‌ها استخراج شد. یک گزارش رویداد دارای ساختار سلسله مراتبی می‌باشد، بدین ترتیب که یک فرایند کسب‌وکار از تعدادی مسیر فرایند تشکیل شده است. هر مسیر فرایند خود شامل چندین رویداد است که هر رویداد حداقل شامل تعدادی ویژگی است. با توجه به این نکته می‌توان دریافت که مشکلات کیفیت داده در سه سطح مختلف مسیر فرایند، رویداد و ویژگی رویداد رخ می‌دهند. همچنین لازم به یادآوری است که ویژگی رویداد حداقل شامل سه خصیصه مورد، فعالیت و برچسب زمانی است. برای هر کدام از سطوح بیان شده، چالش‌های کیفیت داده مربوط به هر کدام از آن‌ها از پژوهش‌های مورد مطالعه استخراج گردیده است که در یک ساختار از بالا به پایین مطابق نمودار (۵) به تصویر کشیده شده‌اند، که در ادامه، به این مشکلات با توجه به سطح مرتبط با آن‌ها، اشاره خواهد شد. همچنین برای هر کدام از این چالش‌ها یک شناسه در نظر گرفته شده است که ترکیبی از ابتدای برگردان انگلیسی واژه مرتبط با آن به همراه یک عدد است.

در جدول‌های شماره (۱ تا ۴) پیوست «الف» به این شناسه‌ها برای مرور پژوهش‌های مورد مطالعه اشاره شده است.

۴-۲-۱. مشکلات کیفیت داده در سطح مسیر فرایند

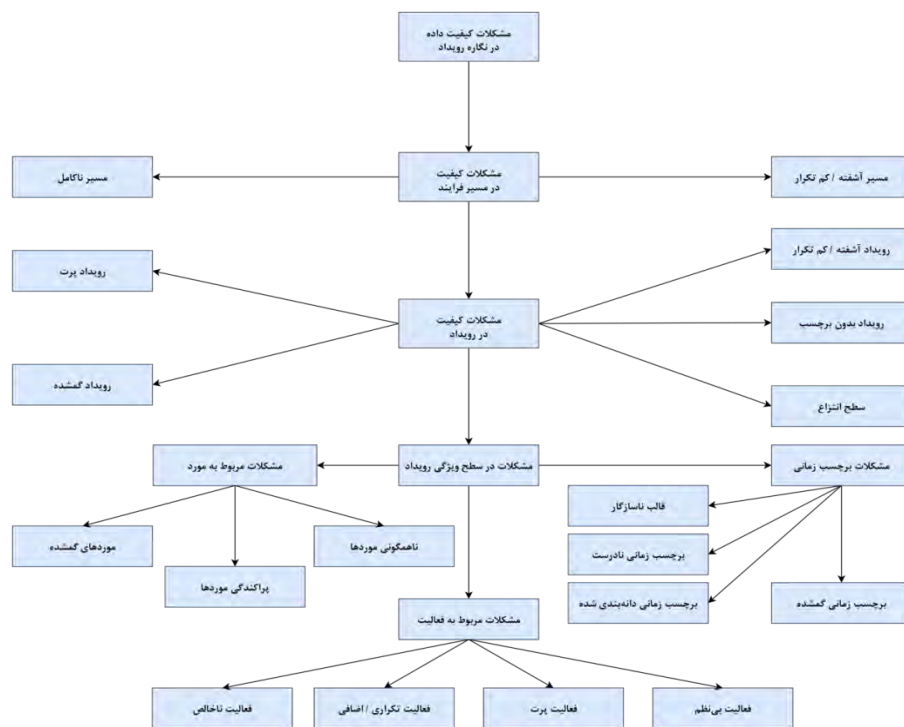
(T1) مسیرهای آشفته^۱/کم تکرار^۲

به شرایطی اشاره دارد که یک یا تعدادی از مسیرهای فرایند، در گزارش رویداد کمتر از دیگر مسیرها اتفاق افتاده‌اند. وجود اینگونه مسیرها منجر به ایجاد ناهنجاری در یک لاگ خواهد شد (Ghionna & et al., 2008).

(T2) مسیر ناکامل^۳

هنگامی که یک فعالیت یا تعدادی از فعالیت‌های موجود، هنگام اجرای یک فرایند موجود نیستند یا ثبت نشده‌اند (Denisov & et al., 2020). در این حالت مسیر فرایند یا به عبارت دیگر دنباله فعالیت‌ها، کامل نخواهد بود.

1. Noise
2. Infrequent
3. Incomplete



نمودار ۵- انواع مشکلات کیفیت داده در گزارش رویداد

۴-۲-۲. مشکلات کیفیت داده در سطح رویداد

(E1) سطح انتزاع: امکان دارد به دلیل سطوح انتزاع مختلف در سامانه‌های اطلاعاتی درگیر در یک فرایند کسب‌وکار، داده‌های رویداد دچار مشکل شوند. به عنوان نمونه یک گزارش رویداد با سطح بالایی از جزئیات، منجر به ایجاد یک مدل فرایند بزرگ و پیچیده شود، حال آنکه یک گزارش رویداد با سطح پایینی از جزئیات، امکان دارد بسیاری از فعالیت‌های کلیدی فرایند را در نظر نگرفته باشد (Bose & et al., 2013) یا به عبارت بهتر، رویداد در لاگ گم شده باشد (Van Der Aalst, 2016).

(E2) رویداد بدون برچسب زمانی: زمانی که یک رویداد در گزارش رویداد ذخیره شده باشد، اما شناسه مورد آن مشخص نباشد، می‌گوییم لاگ دارای رویداد بدون برچسب است (Ferreira & Gillblad, 2009; Helal & Awad, 2020)، یعنی نمی‌توانیم یک رویداد را به یک نمونه فرایند خاص نسبت دهیم.

(E3) نگاشت رویدادها به نمونه فرایندها: گاهی اوقات داده‌های یک رویداد ممکن است در چند جدول ناهمسان ذخیره شده باشند، بنابراین، برای ساخت مسیر (دنباله) برای هر مورد یا نمونه فرایند، باید رویدادهای یک نمونه فرایند به یکدیگر پیوند داده شوند (Song & et al., 2017).

(E4) رویدادهای آشفته/کم تکرار: رویدادهای کم تکرار یا نامتداول، رویدادهایی هستند که خیلی کم در جریان یک فرایند رخ می‌دهند (Bose & et al., 2013). این نوع مشکل بسیار زیاد در پیشینه پژوهش مورد توجه قرار گرفته و در بسیاری از موارد از آن با عنوان «آشفته‌گی» یاد شده است. رویدادهای آشفته یا کم تکرار، منجر به ایجاد تاثیر منفی بر روی خروجی تحلیل‌های فرایندکاوی خواهند شد (Sani, 2020). همچنین در این بخش با نوع دیگری از رفتارها هم روبه‌رو هستیم که با عنوان «بی‌نظم» از آنها یاد می‌شود. رفتارهای بی‌نظم، گونه‌ای از آشفته‌گی هستند که هیچ وابستگی‌ای به اصل فرایند کسب‌وکار ندارند و می‌توانند در هر بخش از فرایند رخ دهند (Yi & Peng, 2019).

(E5) رویدادهای گم‌شده:^۱ زمانی که یک رویداد، به هنگام اجرای یک فرایند کسب‌وکار رخ دهد، اما در گزارش رویداد ذخیره نشود، با مشکل رویداد گم شده روبه‌رو هستیم (Van Der Aalst, 2016).

۴-۲-۳. مشکلات کیفیت داده در سطح ویژگی مورد

(C1) موردهای گم‌شده: شناسه مورد برای یک یا تعدادی از رویدادها وجود ندارد (Van Der Aalst, 2016). این مشکل هم‌ارز مشکل رویداد بدون برچسب است.

(C2) ناهمگونی^۲ موردها: این چالش زمانی پیش می‌آید که به علت تغییر قوانین و شروط در یک فرایند کسب‌وکار، ممکن است رفتارهای اتفاق افتاده برای موردهای ورودی متفاوت شود (Bose & et al., 2013)، به همین علت ممکن است که در یک لاگ، رفتارهای نمونه فرایندها با یکدیگر متفاوت باشد. از این مشکل با عنوان «رانش مفهومی»^۳ نیز یاد می‌شود.

(C3) پراکندگی^۴ موردها: بیانگر شرایطی است که گام‌های مهم و اصلی فرایند در گزارش رویداد موجود نباشند (Suriadi & et al., 2017) و در نتیجه استخراج رویدادهای وابسته به یک مورد (نمونه فرایند) امکان‌پذیر نیست.

1. Missing events
2. Heterogeneity
3. Concept drift
4. Scattered

۴-۲-۴. مشکلات کیفیت داده در سطح ویژگی فعالیت

(A1) **فعالیت‌های ناخالص (آلوده):** فعالیت‌های ذخیره شده برای یک فرایند، بسیار با یکدیگر متفاوت هستند. به عبارت دیگر، برجسب فعالیت‌ها دارای مقادیر پراکنده و ناخالص است (Sadeghianasl & et al., 2019; Suriadi & et al., 2017).

(A2) **فعالیت‌های تکراری:**^۱ به حالتی اشاره دارد که یک فعالیت ممکن است دو مرتبه یا بیش از دو مرتبه، در یک فرایند تکرار شود که در این حالت الگوریتم‌های کشف فرایند ممکن است یک حلقه دور آن فعالیت ایجاد نمایند (Lu & et al., 2016; Lu & et al., 2016)، در صورتی که بنا به دلایلی ممکن است یک فعالیت مشخص چند مرتبه در طی اجرای یک فرایند تکرار شود.

(A3) **فعالیت‌های مترادف:**^۲ هنگامی که برجسب فعالیت‌ها از نظر معنایی با یکدیگر مترادف هستند (Chen & et al., 2021; Sadeghianasl & et al., 2019) و می‌توان آن‌ها را یک فعالیت واحد در نظر گرفت.

(A4) **نگاشت فعالیت‌ها به رویدادها:** به دست آوردن مسیر یک فرایند از طریق نگاشت فعالیت‌های مرتبط با نمونه فرایند به یکدیگر (Baier, 2015).

(A5) **فعالیت‌های بی‌نظم:**^۳ هر فعالیتی که به صورت اتفاقی (بدون برنامه‌ریزی قبلی) در هر نقطه از اجرای فرایند رخ دهد (Tax, Sidorova & et al., 2019).

(A6) **فعالیت‌های گمشده:**^۴ اگر برجسب فعالیتی در مدل مرجع وجود داشته باشد، ولی در گزارش رویداد ثبت نشده باشد، گفته می‌شود که فعالیت در گزارش رویداد گمشده است (Xu & Liu, 2019).

۴-۲-۵. مشکلات در سطح ویژگی برجسب زمانی

(Ti1) **برجسب‌های زمانی نادرست:** زمان ثبت شده به صورت صحیح، زمان و تاریخ یک رویداد را مشخص نمی‌کند، یا زمان و تاریخ به صورت نادرست ذخیره شده است.

(Ti2) **برجسب‌های زمانی دانه‌بندی شده:** منظور تفاوت در سطح انتزاع ذخیره‌سازی زمان و تاریخ یک رویداد است (Bose & et al., 2013) (به عنوان نمونه، یک رویداد هم دارای تاریخ و زمان ثبت

1. Duplicate

2. Synonymous

3. Chaotic

4. Missing Activity

است، اما یک رویداد فقط دارای زمان ثبت است و تاریخ اجرا ندارد).

(Ti3) برجسب‌های زمانی گمشده: زمان ثبت شده برای یک رویداد موجود نیست (Van Der Aalst, 2016).

(Ti4) برجسب‌های زمانی مخلوط: زمان و تاریخ رویدادها ثبت شده است، اما به علت تفاوت در سطح انتزاع داده‌های ذخیره شده، نمی‌توان ترتیب بین رویدادها را تشخیص داد (Bose & et al., 2013).

(Ti5) قالب^۱ زمانی ناسازگار: این مشکل زمانی ایجاد می‌شود که رویدادهای مختلف از قالب‌های زمانی متفاوت، برای ذخیره‌سازی زمان و تاریخ انجام یک رویداد استفاده نمایند (Van Der Aalst, 2011). به همین علت نمی‌توان ترتیب درست رویدادها را شناسایی نمود.

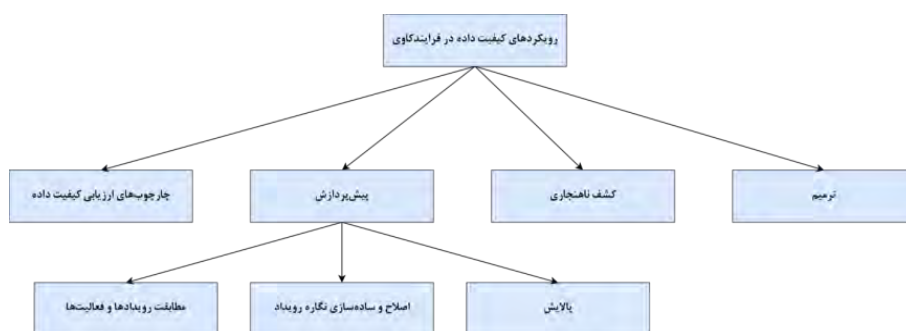
۴-۲-۶. مشکلات کیفیت داده مورد تاکید پژوهش‌ها

بر مبنای بررسی‌های انجام شده، میزان توجه به سه مشکل کیفیت داده در پژوهش‌های مورد بررسی، بیش از سایر موارد بوده است که به ترتیب شامل رویدادهای آشفته / کم تکرار، رویدادهای گمشده و برجسب‌های زمانی نادرست می‌باشند. در رده‌های بعدی مشکلاتی مانند نگاشت رویدادها به نمونه فرایندها، ناهمگونی موردها، رویدادهای بدون برجسب و سطح انتزاع قرار گرفته‌اند. گفتنی است که در پژوهش‌های منتخب، به چالش‌های مربوط به فعالیت‌های گمشده، نگاشت فعالیت‌ها به رویدادها، فعالیت‌های بی‌نظم و قالب زمانی ناسازگار، کمتر از دیگر موارد توجه شده است. نمودار (۶) میزان توجه به انواع مشکلات کیفیت داده در پژوهش‌های بررسی شده را در یک ساختار درختی نشان می‌دهد.



۳-۴. انواع رویکردها برای بررسی و حل چالش کیفیت داده در فرایند کاوی

برای پاسخ به پرسش شماره (۳)، انواع رویکردهای بررسی و حل چالش‌های وابسته به کیفیت داده در فرایند کاوی شناسایی و گزینش شدند. برای این کار در ابتدا اهداف پژوهش‌های مورد مطالعه استخراج شده و سپس مطابق با نمودار (۷) به چهار گروه اصلی تقسیم گردیدند که شامل موارد زیر می‌باشند:



نمودار ۷- رویکردهای پژوهش‌های مربوط به حل مشکلات کیفیت داده در فرایندکاوی

(۱) **چارچوب‌های ارزیابی کیفیت داده:** دسته اول پژوهش‌ها به چارچوب‌های ارزیابی کیفیت داده‌ها اختصاص دارند، که بخش مهمی از پژوهش‌های حاضر در پیشینه به همین گونه تعلق دارند. هدف این نوع پژوهش‌ها بررسی ابعاد اثرگذار بر روی کیفیت گزارش‌های رویداد و ارائه رهیافت‌هایی جامع برای فهم نوع چالش و ویژگی‌های مرتبط با آن‌ها می‌باشد. مطالعات بررسی شده برای این رویکرد در جدول (۱) پیوست «الف» شرح داده شده‌اند.

(۲) **پیش‌پردازش:** دسته دوم پژوهش‌های بررسی شده، به رویکردهای پیش‌پردازش یک گزارش رویداد اشاره دارد. هدف این رویکرد جمع‌آوری و تبدیل داده‌ها به داده‌هایی مناسب برای روش‌های فرایندکاوی، به ویژه کشف فرایند است، تا معیارهای کیفیت نتایج خروجی، بهبود یابند. پژوهش‌های در پیوند با این رویکرد در جدول (۲) پیوست «الف» گردآوری شده‌اند. البته راهبردهای پیش‌پردازش خود به سه زیرگروه پالایش، ساده‌سازی و مطابقت رویدادها و فعالیت‌ها تقسیم می‌گردد:

(۱-۲) **پالایش:**^۱ متداول‌ترین فعالیت پیش‌پردازش است که انگیزه اصلی آن شناسایی و حذف رویدادهای آشفته / کم تکرار یا نادرست از یک گزارش رویداد است تا دقت مدل فرایند کشف شده نهایی افزایش یابد.

(۲-۲) **ساده‌سازی:**^۲ هدف این فعالیت کاهش ابعاد و حجم داده‌های ورودی است تا به صورت همزمان راندمان الگوریتم‌های کشف فرایند بهبود یافته و همچنین دقت مدل کشف شده افزایش یابد.

(۳-۲) **مطابقت رویداد و فعالیت‌ها:** بیشتر مطالعات انجام شده در این حوزه به دنبال کامل کردن مسیرهای فرایند از طریق مطابقت رویدادها با رویدادها و یا مطابقت رویدادها با فعالیت‌ها هستند،

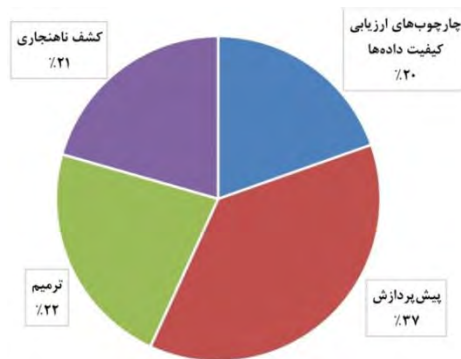
1. Filtering
2. Abstraction

تا بتوانند مسیرهای ناکامل را اصلاح نمایند.

(۳) کشف ناهنجاری:^۱ سومین رویکرد پژوهش‌های بررسی شده، ناظر به تشخیص و حذف رفتارهای ناهنجار ناشی از کیفیت پایین ذخیره سازی داده‌ها می‌باشد. مطالعات بررسی شده در زمینه کشف ناهنجاری در جدول (۳) پیوست «الف» بیان گردیده‌اند.

(۴) ترمیم:^۲ آخرین رویکرد مربوط به پژوهش‌هایی است که هدف آن‌ها بازسازی و بهبود کیفیت گزارش رویداد است. هدف پژوهش‌های این رویکرد اصلاح یا جایگزینی^۳ داده‌های گمشده در گزارش رویداد ورودی است. پژوهش‌های مربوط به این حوزه در جدول (۴) پیوست «الف» توضیح داده شده‌اند.

در نمودار (۸) سهم هر کدام از رویکردهای مورد اشاره در پژوهش‌های مورد بررسی تعیین شده است. براساس نتایج به دست آمده، رویکردهای پیش‌پردازش با ۳۷٪ بیشترین سهم را در پژوهش‌های کیفیت داده در فرایندکاوی به خود اختصاص داده‌اند. در رده‌های بعد مقالات و پژوهش‌های مربوط به چارچوب‌های ارزیابی کیفیت داده‌ها با ۲۰٪، رویکردهای ترمیم و بهبود گزارش رویداد با ۲۲٪ و در انتها رهیافت‌های کشف ناهنجاری با ۲۱٪ قرار گرفته‌اند.



نمودار ۸- رویکردهای پژوهش‌های مورد مطالعه

۴-۴. روش‌های مورد استفاده

برای یافتن پاسخ پرسش (۴)، روش‌ها و رویکردهای مورد استفاده پژوهش‌های مورد بررسی استخراج

1. Anomaly detection
2. Repair
3. Imputation

شده‌اند که گستره پهنآوری از روش‌های ریاضی، آماری و یادگیری ماشین را شامل می‌شوند. براساس بررسی‌های انجام شده، بیشتر مطالعات از «شبکه‌های پتری»^۱ در زمینه کیفیت داده در فرایندکاوی استفاده نموده‌اند. استفاده از این روش در رویکرد پیش‌پردازش و ترمیم، بسیار متداول است. همچنین استفاده از مزایا و ویژگی‌های نظریه اتوماتا^۲ به دلیل قوانین ریاضی مناسب برای مدل‌سازی و تحلیل رفتار پویای موجود در فرایندهای کسب‌وکار، در تعدادی از پژوهش‌ها مورد توجه قرار گرفته است. یکی دیگر از روش‌های محبوب در پژوهش‌های منتخب، روش‌های مبتنی بر نظریه گراف است و در مطالعات زیادی به خصوص در بخش پیش‌پردازش به‌کار گرفته شده‌اند، زیرا به آسانی می‌توان با استفاده از محاسبه تعداد رخداد فعالیت‌ها، فعالیت‌های متداول را از آشفته یا کم تکرار بازشناخت.

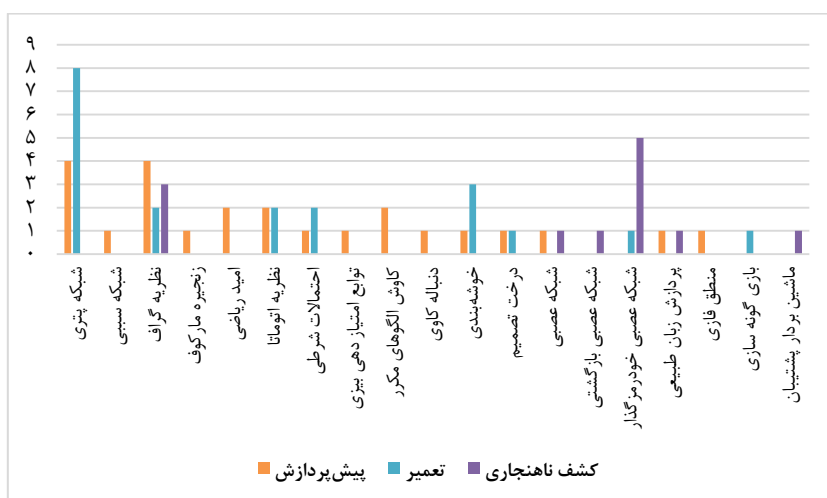
همچنین بهره‌گیری از روش‌های یادگیری بدون نظارت مانند خوشه‌بندی^۳ و کاوش الگوهای مکرر^۴ در فرایندکاوی، به ویژه در زمینه کیفیت داده‌ها، مورد توجه پژوهشگران بوده است؛ زیرا می‌توان به مسیرهای فرایند به عنوان دنباله‌ای از فعالیت‌ها نگریست و بر همین مبنا استفاده از روش‌هایی که می‌توانند ویژگی‌های مکرر را از این دنباله فعالیت‌ها استخراج یا آن‌ها را خوشه‌بندی کنند، بسیار مفید می‌باشد. بیشترین حجم استفاده از این روش‌ها متعلق به رویکردهای پیش‌پردازش و ترمیم گزارش رویداد بوده است.

از روش‌های دسته‌بندی مورد استفاده پژوهش‌ها می‌توان به انواع شبکه‌های عصبی^۵ و درخت تصمیم^۶ اشاره نمود. شبکه‌های عصبی خودرمزگذار^۷ به جهت توانایی یادگیری بدون نظارت، اهمیت زیادی در حوزه فرایندکاوی پیدا کرده‌اند، به ویژه در بخش کشف ناهنجاری که با استفاده از این نوع شبکه، امکان تمایز بین رویدادهای متداول و ناهنجار فراهم خواهد شد. همچنین از روش درخت تصمیم برای پیش‌پردازش و ترمیم رویدادها نیز استفاده شده است.

بهره بردن از رویکردهای احتمالی مانند احتمالات شرطی^۸، زنجیره مارکوف^۹، امید ریاضی^{۱۰} و

1. Petri net
2. Automata theory
3. Clustering
4. Frequent pattern mining
5. Neural networks
6. Decision tree
7. Autoencoder
8. Conditional probability
9. Markov chain
10. Mathematical expectation

احتمالات بیزی، در فرایندکاوی بسیار متداول است؛ زیرا فرایندهای کسب‌وکار مجموعه‌ای از فعالیت‌های قطعی نیستند. به همین جهت استفاده از روش‌های مدل‌سازی و محاسبه احتمالات پیشین^۱ و پسین^۲ بین رویدادها و فعالیت‌ها، برای یافتن فعالیت‌های آشفته یا کم تکرار و همچنین بازسازی مسیر فرایند، مفید خواهد بود. در حال حاضر استفاده از روش‌هایی مانند منطق فازی^۳، بازی‌وارسازی^۴ و ماشین بردار پشتیبان^۵، به صورت اندک در مطالعات مربوط به کیفیت داده در فرایندکاوی مورد توجه بوده است. در نمودار (۹) میزان استفاده از روش‌های مختلف در پژوهش‌های کیفیت داده، در سه رویکرد پیش‌پردازش، ترمیم و کشف ناهنجاری تعیین شده است.



نمودار ۹- مقایسه انواع روش‌های مورد استفاده در پژوهش‌های مرتبط با کیفیت داده در فرایندکاوی

۴-۵. ابزارها و کتابخانه‌های مورد استفاده

در این بخش برای پاسخ به پرسش (۵)، به نتایج حاصل از بررسی چارچوب‌ها و ابزارهای مورد استفاده پژوهش‌های مورد مطالعه خواهیم پرداخت. بر این اساس، سه ابزار مهم که در زمینه کیفیت داده در فرایندکاوی، توسط محققان استفاده شده‌اند، شناسایی گردید. اولین ابزار شناسایی شده،

1. Priori probability
2. Posterior probability
3. Fuzzy logic
4. Gamification
5. Support vector machine

چارچوب ProM^۱ (Verbeek & et al., 2010) می‌باشد که توسط جامعه دانشگاهی و پژوهشگران علاقمند به فرایندکاوی، توسعه و نگهداری می‌شود. ProM یک ابزار رایگان برای استفاده از تکنیک‌های فرایندکاوی است که با استفاده از زبان برنامه‌نویسی جاوا^۲ پیاده‌سازی شده است. دومین ابزار مورد استفاده پژوهش‌های مورد مطالعه، برنامه RapidMiner^۳ است، که یک برنامه تعاملی و گرافیکی برای فعالیت‌های مربوط به علم داده^۴ و یادگیری ماشین می‌باشد. این ابزار از افزونه‌های زیادی بهره می‌برد که یکی از آن‌ها پلاگین RapidProM (van der Aalst & et al., 2017) است که بخشی از توانایی‌های کلیدی ProM را در برنامه RapidMiner قابل استفاده می‌نماید. سومین چارچوب، کتابخانه PM4Py^۵ است که براساس زبان برنامه‌نویسی پایتون توسعه داده شده است و امکان استفاده از روش‌های فرایندکاوی را در محیط برنامه‌نویسی پایتون^۶ فراهم می‌سازد.

در جدول (۵) میزان استفاده از ابزارهای معرفی شده در رویکردهای مورد اشاره در بخش (۴-۳) مشخص شده است. بر این اساس، ابزار ProM بیشترین حجم استفاده را در هر چهار رویکرد به خود اختصاص داده است. در ۳۲ مطالعه مورد بررسی، از ابزار ProM استفاده شده است. در ۷ پژوهش از افزونه RapidProM، و در ۲ مطالعه از کتابخانه PM4Py استفاده شده است.

جدول ۵- انواع چارچوب‌ها و ابزارهای مورد استفاده در پژوهش‌های مورد مطالعه

نوع دسته‌بندی	ProM	RapidProM	PM4Py	مجموع
چارچوب‌های ارزیابی کیفیت داده‌ها	۳	-	-	۳
پیش‌پردازش و پالایش	۱۶	۴	۲	۲۲
بازسازی و تعمیر	۷	۲	-	۹
کشف ناهنجاری	۶	۱	-	۷
مجموع	۳۲	۷	۲	۴۱

۴-۶. مجموعه داده‌های مورد استفاده

به منظور پاسخ به پرسش شماره (۶)، ۱۱ مجموعه داده استاندارد مورد استفاده در پژوهش‌های

1. <https://www.promtools.org/>

2. Java

3. <https://rapidminer.com/>

4. Data science

5. <https://pm4py.fit.fraunhofer.de/>

6. Python

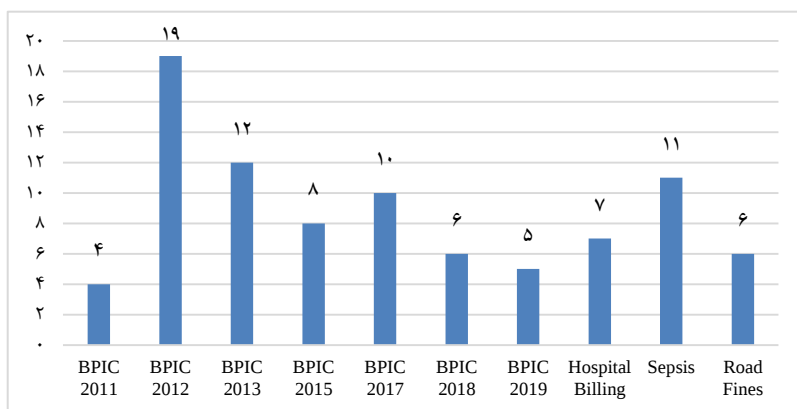
منتخب، گردآوری و دسته‌بندی شدند که نتایج آن در جدول (۳) گزارش شده است. استفاده از مجموعه داده‌های استاندارد به واسطه ایجاد درک مشترک در پژوهشگران و متخصصان، و همچنین به جهت دسترسی آسان، بازاستفاده، بازنشر و مشارکت همگانی، مورد توجه قرار دارند و امکان مقایسه روش‌های مختلف بر مبنای داده‌هایی مشخص و استاندارد را فراهم می‌سازند. بدین‌سان در زمینه فرایندکاوی، چنین مجموعه داده‌هایی برای استفاده پژوهشگران فراهم شده است. در نمودار (۱۰) میزان استفاده از مجموعه داده‌های تشریح شده در جدول (۶)، نشان داده شده است، به صورتی که پژوهش‌های مورد بررسی، به ترتیب بیشتر از مجموعه داده‌های BPIC 2012، BPIC 2013 و Sepsis استفاده نموده‌اند.

جدول ۶- مجموعه داده‌های استاندارد مورد استفاده در پژوهش‌های مورد مطالعه

نام مجموعه داده	زمینه کسب‌وکار	توضیحات	منبع
۱ BPIC 2011	سلامت	داده‌های رویداد مربوط به یک بیمارستان دانشگاهی ^۱ در کشور هلند	(van Dongen, 2011)
۲ BPIC 2012	مؤسسات مالی	گزارش رویدادی شامل رویدادهای مربوط به درخواست وام ^۲	(Van Dongen, 2012)
۳ BPIC 2013	مدیریت سوانح	داده‌های مربوط به یک سامانه مدیریت سوانح و مشکلات ^۳ شرکت Volvo	(Steeman, 2013)
۴ BPIC 2015	شهرداری و خدمات شهری	داده‌های رویداد مربوط به شهرداری‌های هلند	(van Dongen, 2015)
۵ BPIC 2017	مؤسسات مالی	مجموعه رویدادهایی مربوط به درخواست وام از یک موسسه مالی هلندی ^۴	(van Dongen, 2017)
۶ BPIC 2018	صنعت کشاورزی	داده‌های رویداد مربوط به یک سامانه اطلاعاتی مربوط به سیاست‌های کشاورزی مشترک اتحادیه اروپا ^۵	(van Dongen, 2018)
۷ BPIC 2019	صنعت نساجی	داده‌های متعلق به یک شرکت بزرگ چند ملیتی فعال در حوزه صنایع نساجی و رنگ	(van Dongen, 2019)

1. Academic hospital
2. Loan application
3. Incident and problem management system
4. Dutch financial institute
5. Common Agricultural Policy

نام مجموعه داده	زمینه کسب‌وکار	توضیحات	منبع
۸ Hospital Billing	سلامت	داده‌های رویداد بخش مالی یک سامانه ERP در یک بیمارستان منطقه‌ای	(Mannhardt, 2017)
۹ Sepsis	سلامت	داده‌های رویداد برای درمان بیماری «سپسیس» ^۱ که از یک سامانه ERP جمع‌آوری شده‌اند	(Mannhardt, 2016)
۱۰ Road Fines	صنعت حمل و نقل	داده‌های رویداد یک سامانه مدیریت جرایم ترافیک جاده‌ای	(De Leoni & Felix, 2015)



نمودار ۱۰- میزان استفاده از مجموعه داده‌های استاندارد در پژوهش‌های کیفیت داده در فرایندکاوی

۵. نتیجه‌گیری

مطالعه حاضر نشان داد که با افزایش استفاده از روش‌های فرایندکاوی، اهمیت توجه به کیفیت داده‌های ورودی برای شناسایی و بهبود فرایندهای کسب‌وکار افزایش یافته است. روند زمانی مطالعات انجام شده نشان داد که به صورت میانگین بیش از ۶۰ درصد پژوهش‌های مربوط به این زمینه بین سال‌های ۲۰۱۸ تا ۲۰۲۱ میلادی انجام شده است که فرض ما در مورد افزایش توجه به موضوع کیفیت داده در فرایندکاوی در سال‌های اخیر را تایید می‌نماید. همچنین نتایج ارزیابی‌ها نشان داد که مطالعات مورد بررسی متعلق به بیش از ۴۰ همایش و ۳۰ مجله معتبر بوده و از نظر تعداد به ترتیب توسط سه انتشارات Springer، IEEE و Elsevier نمایه و چاپ شده‌اند.

بررسی مطالعات انجام شده نشان داد که ممکن است ۲۰ نوع مشکل کیفیت داده در یک گزارش

رویداد ایجاد شود که در این پژوهش، در ۵ سطح مسیر فرایند، رویداد، مورد، فعالیت و برجسب زمانی دسته‌بندی گردیدند. مطابق بررسی‌های انجام شده، چالش‌های مربوط به رویدادهای کم تکرار، پرت و آشفته، بیشترین توجهات را در مطالعات کیفیت داده در فرایندکاوی به خود اختصاص داده‌اند. همچنین برای ایجاد دیدگاهی بهتر، مطالعات منتخب بر مبنای هدف آن‌ها، به چهار دسته اصلی تقسیم شدند. بر این اساس، اولین رویکردهایی که در زمینه کیفیت داده در فرایندکاوی به مرحله نشر رسیده‌اند، چارچوب‌های ارزیابی کیفیت داده در یک گزارش رویداد بوده‌اند. در این دسته ۲۰ پژوهش مهم شناسایی شدند که از این تعداد ۱۱ مورد پژوهش‌هایی با رویکرد عمومی بوده‌اند که در بیشتر سازمان‌ها یا کسب‌وکارها قابل استفاده هستند و ۹ پژوهش دیگر برای زمینه‌های خاص توسعه داده شده بودند. بیشترین حوزه مورد بررسی ادبیات در این دسته، مربوط به بخش سلامت است. به عنوان شکاف‌های پژوهش در این گروه می‌توان به توسعه چارچوب‌های کیفیت داده عمومی بر مبنای استانداردها و ابعاد کیفیت داده اشاره نمود. همچنین توسعه چارچوب‌هایی برای ارزیابی کیفیت داده در دیگر زمینه‌های کسب‌وکار از جمله بخش‌های مالی و بانکی، بیمه، آموزش، شرکت‌های خدماتی، صنایع مربوط به حوزه حمل و نقل که از کاربران استفاده‌کننده از فرایندکاوی هستند، هم می‌تواند با جدیت بیشتری مورد توجه قرار گیرد.

دسته دوم رویکردهای شناسایی شده، پیش‌پردازش است که ۳۸ مطالعه در این گروه استقرار یافته‌اند. پیش‌پردازش خود به سه زیرگروه پالایش، ساده‌سازی و مطابقت رویدادها و فعالیت‌ها تقسیم می‌شود. پژوهش‌های مربوط به بخش پالایش ۲۰ مورد بوده که انگیزه آن‌ها بیشتر به حذف داده‌های آشفته و کم تکرار از گزارش رویداد معطوف بوده است. از طرفی چون روش‌های فرایندکاوی در مواجهه با داده‌های ورودی، دارای حجم زیاد یا پیچیدگی بالا دچار چالش خواهند شد، ساده‌سازی گزارش رویداد یک راهبرد مناسب برای کاهش ابعاد و پیچیدگی داده‌های ورودی خواهد بود که ۱۱ مطالعه به این بخش تعلق دارند. همچنین یافتن همخوانی بین رویدادها و فعالیت‌ها که گامی حیاتی برای ساخت مسیر فرایند محسوب می‌شود، در ۷ پژوهش مورد بررسی و تحلیل قرار گرفته است. در این گروه یافتن حد آستانه بهینه برای تمایز بین رویدادهای آشفته یا کم تکرار، با رویدادهای متداول به همراه توسعه روش‌هایی که امکان یافتن همخوانی بهینه بین رویدادهای پیچیده را داشته باشند، که دارای روابط موازی و همزمان زیادی هستند، همچنان به عنوان چالشی باز در فرایندکاوی به‌شمار می‌رود.

دسته سوم شامل روش‌های کشف ناهنجاری در گزارش رویداد با زمینه کیفیت داده است. پژوهش‌های مربوط به این دسته ۲۰ مورد بودند که بررسی پژوهش‌های سال‌های اخیر نشان می‌دهد توجه پژوهشگران به این گروه بسیار افزایش یافته است. یکی از چالش‌های مهم این حوزه، ارائه

روش‌هایی بدون نظارت برای طبقه‌بندی رویدادها به رویدادهای متداول و ناهنجار است؛ زیرا امکان دارد در بعضی از شرایط، داده‌های ورودی بدون برچسب بوده، یا مدل فرایند مرجع موجود نباشد. بدین گونه، بخش زیادی از این مطالعات تلاش کرده‌اند از روش‌های یادگیری بدون نظارت بهره ببرند و در نتیجه به‌کارگیری شبکه‌های عصبی خودرمزگذار، در این دسته از مطالعات مورد تاکید بوده است. البته باید گفت که شناخت رویدادهای ناهنجار، به آسانی امکان‌پذیر نخواهد بود، زیرا در یک گزارش رویداد پیچیده که فعالیت‌های دارای وابستگی زیاد یا متوازی فراوان هستند، روش‌های ارائه شده به جهت یافتن مقدار آستانه بهینه و همچنین زمان اجرا، دچار چالش خواهند شد. به همین جهت استفاده از پیشرفت‌های اخیر در حوزه یادگیری عمیق می‌تواند کمک شایانی به یادگیری لایه‌های پنهان داده‌ها نموده، تا امکان یافتن خودکار مقادیر آستانه مناسب برای تمایز بین رویدادهای متداول و نامناسب را فراهم سازد.

آخرین رویکرد شناسایی شده مربوط به بهبود کیفیت یک گزارش رویداد با استفاده از رویکردهای بازسازی و تعمیر است. ۲۴ پژوهش مورد بررسی در این دسته قرار گرفتند. چالشی که رویکردهای این گروه به دنبال چاره‌اندیشی آن بوده‌اند، توسعه روش‌هایی برای بازیابی رویدادهای گمشده، به ویژه در چالش‌های کیفیت داده‌ای چون شناسه نمونه فرایندهای گمشده یا فعالیت‌های گمشده بوده است. همچنین شناسایی برچسب‌های زمانی گمشده و بازسازی مجدد ترتیب رویدادها براساس جایگزینی مقادیر برچسب زمانی مورد توجه بوده است. با توجه به بزرگی و پیچیدگی فرایندهای کسب‌وکار و احتمال رخداد حالت‌های ویژه یا تغییرپذیری فرایند، با توجه به شروط کسب‌وکار، فضای پاسخ برای یافتن مقدار گمشده یا نزدیک به آن، بسیار بزرگ خواهد بود. در نتیجه بیشتر مطالعات این دسته از روش‌های اکتشافی^۱ برای کاهش فضای جستجو استفاده نموده‌اند. از این‌رو امکان دارد روش‌های جاری به هنگام مواجه شدن با فرایندهای پیچیده و دارای مقادیر گمشده، زیاد دچار مشکل شوند که به عنوان یک شکاف پژوهشی درخور نگرش است. همچنین استفاده از رویکردهای نوین کاهش ابعاد که به شکل گسترده‌ای در مطالعات مربوط به یادگیری ماشین و داده‌کاوی مورد استفاده قرار می‌گیرند، می‌تواند مورد بهره‌برداری قرار گیرد. از طرفی توجه به دیگر ویژگی‌های گمشده، برای نمونه منابع گمشده (شامل افراد یا منابع سازمانی) و وابستگی آن‌ها به فعالیت‌های کلیدی می‌تواند رهیافتی نوین برای بازسازی مقادیر گمشده باشد.

همچنین بررسی پژوهش‌های منتخب نشان داد که بیشتر آن‌ها از ابزارهای ProM و RapidProM و کتابخانه PM4Py برای پیاده‌سازی روش‌های ارائه شده استفاده نموده‌اند که استفاده از ابزار ProM بیشتر از موارد دیگر بوده است. علاوه بر این چون فرایندکاوی به عنوان یکی از زیرشاخه‌های علم داده شناخته می‌شود، ساخت مجموعه داده‌های استاندارد برای ارزیابی و مقایسه نتایج حاصل از پژوهش‌های فرایندکاوی از اهمیت زیادی برخوردار شده است. نتایج این مطالعه نشان داد که ۸۸ پژوهش از مجموعه ۱۰۲ پژوهش مورد مطالعه، از مجموعه داده‌های استاندارد برای ارزیابی و تحلیل نتایج خود استفاده نموده‌اند.

اجرای این مطالعه با دو چالش مهم روبه‌رو بوده است. نخست اینکه بعضی از پژوهش‌ها در پایگاه‌های عملی یا نشریاتی منتشر شده بودند که دسترسی به آن‌ها فراهم نگردید. از طرفی، فقط مطالعاتی مورد کاوش قرار گرفته‌اند که به زبان انگلیسی نوشته شده باشند، زیرا هیچ منبع نگارش یافته به زبان فارسی که به زمینه کیفیت داده در فرایندکاوی پرداخته باشد، یافت نشد. سخن پایانی اینکه، مطالعه حاضر که از روش مرور نظام‌مند استفاده نموده، کوششی است برای بازگویی مشکلات کیفیت داده در فرایندکاوی و دسته‌بندی رویکردهای ارائه شده برای حل این چالش‌ها که می‌تواند توسط دیگر پژوهشگران و علاقمندان به زمینه فرایندکاوی مورد استفاده قرار گیرد.

منابع

- حسینی، س.، مصلح، ع.، حسینی، م. (۱۳۹۷). تحلیل فرایندهای الکترونیکی با استفاده از تکنیک فرایندکاوی (مورد مطالعه: فرایند ترفیع پایه اعضای هیئت علمی دانشگاه خلیج فارس). چشم‌انداز مدیریت صنعتی، ۸(۲۹): ۱۱۳-۱۳۵.
- خجسته، ف.، کاهانی، م.، بهکمال، ب. (۱۳۹۹). شناسایی رانش مفهومی در نگاره‌های فرایند کسب‌وکار با استفاده از یادگیری عمیق. پردازش‌ها، ۱۷(۴): ۳۳-۴۸.
- خدیور، آ.، فیروزی، ف.، نیاکان، ل. (۱۳۹۹). ارزیابی ریسک و بررسی انطباق با قواعد کسب وکار در بیمه با کمک تکنیک فرایندکاوی مطالعه موردی: فرایند خسارت بیمه شخص ثالث. پژوهشنامه بیمه (صنعت بیمه)، ۳۵(۲): ۳۳-۶۵.
- خوش‌خوی‌نیلاش، ا.، تمجیدیامچلو، ع.، راد، ر. (۱۴۰۰). تحلیل عملکرد و بهبود فرایندهای ارائه تسهیلات سرمایه در گردش بانک صنعت و معدن با رویکرد فرایندکاوی. مطالعات مدیریت کسب وکار هوشمند، ۳۹(۱): ۷۰-۷۰.
- شامی زنجانی، م.، نبی، ف.، ایران دوست، ش. (۱۳۹۹). ناخدایی دیجیتال: راهنمای تحول سازمان‌ها در عصر دیجیتال. تهران: انتشارات آریانا قلم.
- علیزاده، ف.، هادوی‌نژاد، م. (۱۳۹۸). فرایندکاوی نخبه‌ستیزی در سازمان براساس طرح نگاشت تفسیری نظریه داده‌بنیاد. پژوهش‌های مدیریت منابع سازمانی، ۹(۱): ۱۶۵-۱۸۳.
- مصطفایی دولت‌آباد، خ.، آذر، ع.، مقبل باعرض، ع.، پرویزیان، ک. (۱۳۹۸). ارزیابی فرایندکاوی در کشف مدل فرایندهای نیمه اتوماتیک صنعت بانکداری (مورد مطالعه فرایند صدور ضمانت‌نامه بانکی). مطالعات مدیریت صنعتی، ۱۷(۵۲): ۳۷-۱.
- ون در آلست، و. (۱۳۹۴). فرایندکاوی: کشف، تطبیق و بهبود فرایندهای کسب‌وکار. ترجمه س.ح. سیادت و ر. همتی گشتاسب. تهران: دانشگاه شهید بهشتی.

References

- Alharbi, A., Bulpitt, A. & Johnson, O. (2017). *Improving pattern detection in healthcare process mining using an interval-based event selection method*. Business Process Management Forum: BPM Forum 2017, Barcelona, Spain, September 10-15. Proceedings 15.
- Al-Hashedi, K.G. & Magalingam, P. (2021). Financial fraud detection applying data mining techniques: a comprehensive review from 2009 to 2019. *Computer Science Review*, 40: 100402. <https://doi.org/10.1016/j.cosrev.2021.100402>
- Alizadeh, F. & Hadavinejad, M. (2019). Process Mining for Anti-Elitism in an Organization based on Interpretive Mapping Design of Grounded Theory. *Organizational Resources Management Researchs*, 9(1): 165-183. [in persian]
- Andrews, R., Emamjome, F., ter Hofstede, A.H. & Reijers, H.A. (2020). *An expert lens on data quality in process mining*. 2020 2nd International Conference on Process Mining (ICPM).
- Andrews, R., Suriadi, S., Ouyang, C. & Poppe, E. (2018). *Towards event log querying for data quality: Let's start with detecting log imperfections*. On the Move to Meaningful Internet Systems. OTM 2018 Conferences: Confederated International Conferences: CoopIS, C&TC, and ODBASE 2018, Valletta, Malta, October 22-26. Proceedings, Part I.
- Andrews, R., van Dun, C.G., Wynn, M.T., Kratsch, W., Röglinger, M. & ter Hofstede, A.H. (2020). Quality-informed semi-automated event log generation for process mining. *Decision*

- Support Systems*, 132: 113265. <https://doi.org/10.1016/j.dss.2020.113265>
- Andrews, R., Wynn, M.T., Vallmuur, K., Ter Hofstede, A.H., Bosley, E., Elcock, M. & Rashford, S. (2019). Leveraging data quality to better prepare for process mining: an approach illustrated through analysing road trauma pre-hospital retrieval and transport processes in Queensland. *International journal of environmental research and public health*, 16(7): 1138. <https://doi.org/10.3390/ijerph16071138>
- Ayo, F.E., Folorunso, O. & Ibhara, F.T. (2017). A probabilistic approach to event log completeness. *Expert Systems with Applications*, 80: 263-272. <https://doi.org/10.1016/j.eswa.2017.03.039>
- Baier, T. (2015). *Matching events and activities: preprocessing event logs for process analysis*. Doctoral dissertation. Universität Potsdam. Retrieved from: <https://publishup.uni-potsdam.de/frontdoor/index/index/docId/8454>
- Bayomie, D., Awad, A. & Ezat, E. (2016). *Correlating unlabeled events from cyclic business processes execution*. International Conference on Advanced Information Systems Engineering.
- Bernard, G. & Andritsos, P. (2020). *Truncated trace classifier. removal of incomplete traces from event logs*. Enterprise, Business-Process and Information Systems Modeling: 21st International Conference, BPMDS 2020, 25th International Conference, EMMSAD 2020, Held at CAiSE 2020, Grenoble, France, June 8–9, Proceedings 21.
- Bezerra, F. & Wainer, J. (2013). Algorithms for anomaly detection of traces in logs of process aware information systems. *Information systems*, 38(1): 33-44. <https://doi.org/10.1016/j.is.2012.04.004>
- Bezerra, F., Wainer, J. & van der Aalst, W.M. (2009). *Anomaly detection using process mining*. Enterprise, Business-Process and Information Systems Modeling: 10th International Workshop, BPMDS 2009, and 14th International Conference, EMMSAD 2009, held at CAiSE 2009, Amsterdam, The Netherlands, June 8-9. Proceedings.
- Böhmer, K. & Rinderle-Ma, S. (2016). *Multi-perspective anomaly detection in business process execution events*. On the Move to Meaningful Internet Systems: OTM 2016 Conferences: Confederated International Conferences: CoopIS, C & TC, and ODBASE 2016, Rhodes, Greece, October 24-28, Proceedings.
- Bose, J.C. (2012). *Process mining in the large: Preprocessing, Discovery and Diagnostics*. Eindhoven University of Technology. Eindhoven.
- Bose, R.P.J.C., Mans, R.S. & van der Aalst, W.M. (2013). *Wanna improve process mining results?: it's high time we consider data quality issues seriously*. BPM reports, 1302.
- Burattin, A. (2015). *Process mining techniques in business environments*. Springer Cham. <https://doi.org/10.1007/978-3-319-17482-2>
- Cappiello, C., Comuzzi, M., Plebani, P. & Fim, M. (2022). Assessing and improving measurability of process performance indicators based on quality of logs. *Information systems*, 103: 101874. <https://doi.org/10.1016/j.is.2021.101874>
- Chapela-Campa, D., Mucientes, M. & Lama, M. (2019). *Simplification of complex process models by abstracting infrequent behaviour*. Service-Oriented Computing: 17th International Conference, ICSOC 2019, Toulouse, France, October 28–31, 2019, Proceedings 17.

- Chen, L., Kang, S., Karimidorabati, S. & Haas, C. (2019). *Improving the quality of event logs in the construction industry for process mining*. ISARC. Proceedings of the International Symposium on Automation and Robotics in Construction.
- Chen, Q., Lu, Y., Tam, C. & Poon, S. (2021). A Novel Approach to Detect Redundant Activity Labels for More Representative Event Logs. *arXiv preprint arXiv*: 2103.16061.
- Cheng, H.-J. & Kumar, A. (2015). Process mining on noisy logs - Can log sanitization help to improve performance? *Decision Support Systems*, 79: 138-149.
<https://doi.org/10.1016/j.dss.2015.08.003>
- Commission, I.O.F.S.I.E. (2008). Software engineering-Software product Quality Requirements and Evaluation (SQuaRe) Data quality model. *ISO/IEC*, 25012: 1-13.
- Conforti, R., La Rosa, M. & ter Hofstede, A.H. (2016). Filtering out infrequent behavior from business process event logs. *IEEE Transactions on Knowledge and Data Engineering*, 29(2): 300-314. <https://doi.org/10.1109/TKDE.2016.2614680>
- Conforti, R., La Rosa, M., Ter Hofstede, A.H. & Augusto, A. (2020). *Automatic repair of same-timestamp errors in business process event logs*. Business Process Management: 18th International Conference, BPM 2020, Seville, Spain, September 13–18, 2020, Proceedings 18.
- Conforti, R., Rosa, M.L. & Hofstede, A.H.M.T. (2018). *Timestamp Repair for Business Process Event Logs*. University of Melbourne: Melbourne, Australia.
- De Leoni, M. & Felix, M. (2015). *Road Traffic Fine Management Process*. 4TU.ResearchData. Dataset. <https://doi.org/10.4121/uuid:270fd440-1057-4fb9-89a9-b699b47990f5>
- Denisov, V., Fahland, D. & van der Aalst, W.M. (2020). *Repairing event logs with missing events to support performance analysis of systems with shared resources*. Application and Theory of Petri Nets and Concurrency: 41st International Conference, PETRI NETS 2020, Paris, France, June 24–25, 2020, Proceedings 41.
- Díaz-Rodríguez, O.E. & Hernández, M.G.P. (2020). *Quality event log to intention mining: a study case*. In: 2020 International Conference on Computer Science, Engineering and Applications (ICCSEA).
- Dixit, P.M., Suriadi, S., Andrews, R., Wynn, M.T., ter Hofstede, A.H., Buijs, J.C. & van der Aalst, W.M. (2018). *Detection and interactive repair of event ordering imperfection in process logs*. Advanced Information Systems Engineering: 30th International Conference, CAiSE 2018, Tallinn, Estonia, June 11-15, 2018, Proceedings 30.
- Dumas, M., La Rosa, M., Mendling, J. & Reijers, H.A. (2018). *Fundamentals of Business Process Management*. (2 ed.). Springer Berlin, Heidelberg. <https://doi.org/10.1007/978-3-662-56509-4>
- Emanjome, F., Andrews, R., ter Hofstede, A. & Reijers, H. (2020). *Alohomora: Unlocking data quality causes through event log context*. Proceedings of the 28th European Conference on Information Systems (ECIS2020).
- Erdogan, T.G. & Tarhan, A. (2018). Systematic mapping of process mining studies in healthcare. *IEEE Access*, 6: 24543-24567. <https://doi.org/10.1109/ACCESS.2018.2831244>
- Fan, W. (2012). *Data Quality: Theory and Practice*. Web-Age Information Management 13th International Conference, WAIM 2012, Harbin, China.
- Ferreira, D.R. & Gillblad, D. (2009). *Discovering process models from unlabelled event logs*. Business Process Management: 7th International Conference, BPM 2009, Ulm, Germany,

- September 8-10, 2009. Proceedings 7.
- Fischer, D.A., Goel, K., Andrews, R., van Dun, C.G.J., Wynn, M.T. & Röglinger, M. (2020). *Enhancing event log quality: Detecting and quantifying timestamp imperfections*. Business Process Management: 18th International Conference, BPM 2020, Seville, Spain, September 13–18, 2020, Proceedings 18.
- Fox, F., Aggarwal, V.R., Whelton, H. & Johnson, O. (2018). *A data quality framework for process mining of electronic health record data*. 2018 IEEE International Conference on Healthcare Informatics (ICHI).
- Gao, Y., Song, S., Zhu, X., Wang, J., Lian, X. & Zou, L. (2018). Matching heterogeneous event data. *IEEE Transactions on Knowledge and Data Engineering*, 30(11): 2157-2170. <https://doi.org/10.1109/TKDE.2018.2815695>
- Ghionna, L., Greco, G., Guzzo, A. & Pontieri, L. (2008). *Outlier detection techniques for process mining applications*. Foundations of Intelligent Systems: 17th International Symposium, ISMIS 2008 Toronto, Canada, May 20-23, 2008 Proceedings 17.
- Helal, I. & Awad, A. (2020). Correlating Unlabeled Events at Runtime. *arXiv preprint arXiv*, 2004.09971. <https://doi.org/10.48550/arXiv.2004.09971>
- Horita, H., Kurihashi, Y. & Miyamori, N. (2020). Extraction of missing tendency using decision tree learning in business process event log. *Data*, 5(3): 82. <https://doi.org/10.3390/data5030082>
- Hosseini, S., Mosleh, A. & Hosseini, M. (2018). Analyzing Organizational Processes by using Process Mining Technique (The Case of Academic Staff's Grade Process of Persian Gulf University). *Industrial Management Perspective*, 8(1): 113-135. [in persian]
- Huang, Y., Wang, Y. & Huang, Y. (2018). *Filtering Out Infrequent Events by Expectation from Business Process Event Logs*. 2018 14th International Conference on Computational Intelligence and Security (CIS).
- Huang, Y., Zhong, L. & Chen, Y. (2020). Filtering Infrequent Behavior in Business Process Discovery by Using the Minimum Expectation. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 14(2): 1-15.
- Huo, S., Völzer, H., Reddy, P., Agarwal, P., Isahagian, V. & Muthusamy, V. (2021). *Graph autoencoders for business process anomaly detection*. Business Process Management: 19th International Conference, BPM 2021, Rome, Italy, September 06–10, 2021, Proceedings 19.
- Ingvaldsen, J.E. & Gulla, J.A. (2007). *Preprocessing support for large scale process mining of SAP transactions*. International Conference on Business process management.
- Junior, S.B., Ceravolo, P., Damiani, E., Omori, N.J. & Tavares, G.M. (2020). *Anomaly detection on event logs with a scarcity of labels*. 2020 2nd International Conference on Process Mining (ICPM).
- Khadivar, A., Frouzi, F. & Niakan, L. (2020). Risk Analysis and Compliancechecking of Business Rules in Insurance by using of Process Mining Technique. *Iranian Journal of Insurance Research (IJIR)*, 9(2): 33-65. <https://doi.org/10.22056/ijir.2020.02.02> [in persian]
- Kherbouche, M.O., Laga, N. & Masse, P.-A. (2016). *Towards a better assessment of event logs quality*. 2016 IEEE Symposium Series on Computational Intelligence (SSCI).
- Khojasteh, F., Kahani, M. & Behkamal, B. (2021). Concept Drift Detection in Business Process

- Logs using Deep Learning. *Signal and Data Processing*, 17(4): 33-48.
<http://dx.doi.org/10.29252/jsdp.17.4.33> [in persian]
- Khoshkhouy Nilash, E., Tamjid Yamechlo, A. & Rad, R. (2021). Performance Analysis and Improvement of Bank of Industry and Mine Working Capital Facility Processes Based on Process Mining Approach. *Business Intelligence Management Studies*, 9(36): 37-71.
<https://doi.org/10.22054/ims.2021.58106.1896> [in persian]
- Ko, J. & Comuzzi, M. (2020). *Online anomaly detection using statistical leverage for streaming business process events*. International Conference on Process Mining.
- Ko, J. & Comuzzi, M. (2021). *Online anomaly detection using statistical leverage for streaming business process events*. Process Mining Workshops: ICPM 2020 International Workshops, Padua, Italy, October 5–8, 2020, Revised Selected Papers 2.
- Kong, L., Li, C., Ge, J., Li, Z., Zhang, F. & Luo, B. (2019). *An Efficient Heuristic Method for Repairing Event Logs Independent of Process Models*. IoTBDS.
- Krajsic, P. & Franczyk, B. (2020). *Lambda Architecture for Anomaly Detection in Online Process Mining Using Autoencoders*. International Conference on Computational Collective Intelligence.
- Krajsic, P. & Franczyk, B. (2021). *Variational Autoencoder for Anomaly Detection in Event Data in Online Process Mining*. ICEIS.
- Kurniati, A.P., Rojas, E., Hogg, D., Hall, G. & Johnson, O.A. (2019). The assessment of data quality issues for process mining in healthcare using Medical Information Mart for Intensive Care III, a freely available e-health record database. *Health informatics journal*, 25(4): 1878-1893.
<https://doi.org/10.1177/1460458218810760>
- Laranjeiro, N., Soydemir, S.N. & Bernardino, J. (2015). *A survey on data quality: classifying poor data*. 2015 IEEE 21st Pacific rim international symposium on dependable computing (PRDC).
- Leemans, S.J., Fahland, D. & van der Aalst, W.M. (2013). *Discovering block-structured process models from event logs containing infrequent behaviour*. International conference on business process management.
- Li, F. (2020). Leading digital transformation: three emerging approaches for managing the transition. *International Journal of Operations & Production Management*, 40(6): 809-817.
<https://doi.org/10.1108/IJOPM-04-2020-0202>
- Liu, J., Xu, J., Zhang, R. & Reiff-Marganiec, S. (2021). A repairing missing activities approach with succession relation for event logs. *Knowledge and Information Systems*, 63(2): 477-495.
<https://doi.org/10.1007/s10115-020-01524-6>
- Lu, K., Fang, X., Fang, N. & Asare, E. (2021). Discovery of effective infrequent sequences based on maximum probability path. *Connection Science*, 34(1): 63-82.
<https://doi.org/10.1080/09540091.2021.1951667>
- Lu, X. & Fahland, D. (2017). *A Conceptual Framework for Understanding Event Data Quality for Behavior Analysis*. 9th Central European Workshop on Services and their Composition Zeus Workshop 2017, 13-14 februari 2017, Lugano, Switzerland, Switzerland.
- Lu, X. (2016). *Handling Duplicated Tasks in Process Discovery by Refining Event Labels*. Business Process Management: 14th International Conference, BPM 2016, Rio de Janeiro, Brazil, September 18-22. Proceedings 14. Springer International Publishing.

- <https://doi.org/10.4121/uuid:ea90c4be-64b6-4f4b-b27c-10ede28da6b6>
- Lu, X., Fahland, D. & van der Aalst, W.M. (b2016). *Interactively Exploring Logs and Mining Models with Clustering, Filtering, and Relabeling*. Proceedings of the BPM 2016 Tool Demonstration Track.
- Lu, X., Fahland, D., Andrews, R., Suriadi, S., Wynn, M.T., ter Hofstede, A.H. & van der Aalst, W.M. (2017). *Semi-supervised log pattern detection and exploration using event concurrence and contextual information*. On the Move to Meaningful Internet Systems. OTM 2017 Conferences: Confederated International Conferences: CoopIS, C&TC, and ODBASE 2017, Rhodes, Greece, October 23-27, 2017, Proceedings, Part I.
- Lu, X., Fahland, D., van den Biggelaar, F.J. & van der Aalst, W.M. (a2016). *Handling duplicated tasks in process discovery by refining event labels*. Business Process Management: 14th International Conference, BPM 2016, Rio de Janeiro, Brazil, September 18-22, 2016. Proceedings 14.
- Ly, L.T., Indiono, C., Mangler, J. & Rinderle-Ma, S. (2012). *Data transformation and semantic log purging for process mining*. International Conference on Advanced Information Systems Engineering.
- Mannhardt, F. (2016). *Sepsis Cases - Event Log*. 4TU. ResearchData. Dataset.
<https://doi.org/10.4121/uuid:915d2bfb-7e84-49ad-a286-dc35f063a460>
- Mannhardt, F. (2017). *Hospital Billing - Event Log*. 4TU.ResearchData. Dataset.
<https://doi.org/10.4121/uuid:76c46b83-c930-4798-a1c9-4be94dfef741>
- Martin, N. (2018). *Using indoor location system data to enhance the quality of healthcare event logs: opportunities and challenges*. International Conference on Business Process Management.
- Martin, N. (2019). *Using indoor location system data to enhance the quality of healthcare event logs: opportunities and challenges*. Business Process Management Workshops: BPM 2018 International Workshops, Sydney, NSW, Australia, September 9-14, 2018, Revised Papers 16.
- Martin, N. (2021). *Data Quality in Process Mining*. Springer.
https://doi.org/10.1007/978-3-030-53993-1_5
- Mingers, J. & Willcocks, L. (2014). An integrative semiotic framework for information systems: The social, personal and material worlds. *Information and Organization*, 24(1): 48-70.
<https://doi.org/10.1016/j.infoandorg.2014.01.002>
- Mostafaei Dolatabad, K., Azar, A., Moghbel Baarz, A. & Parvizian, K. (2019). Mining Process Evaluation in Discovering the Semi-Automatic Processes of the Banking Industry (the case: Bank guarantee issuance process). *Industrial Management Studies*, 17(52): 1-37. [in persian]
- Mueller-Wickop, N. & Schultz, M. (2013). *ERP event log preprocessing: timestamps vs. accounting logic*. Design Science at the Intersection of Physical and Virtual Design: 8th International Conference, DESRIST 2013, Helsinki, Finland, June 11-12, 2013. Proceedings 8.
- Ngai, E.W., Hu, Y., Wong, Y.H., Chen, Y. & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, 50(3): 559-569. <https://doi.org/10.1016/j.dss.2010.08.006>
- Nguyen, H.T.C. & Comuzzi, M. (2018). *Event Log Reconstruction Using Autoencoders*. International Conference on Service-Oriented Computing.
- Nguyen, H.T.C. (2018). *Neural Computing for Event Log Quality Improvement* Graduate School of

- UNIST]. URL= <http://unist.dcollection.net/common/orgView/200000011233>
- Nguyen, H.T.C., Lee, S., Kim, J., Ko, J. & Comuzzi, M. (2019). Autoencoders for improving quality of process event logs. *Expert Systems with Applications*, 131: 132-147. <https://doi.org/10.1016/j.eswa.2019.04.052>
- Nolle, T., Luettggen, S., Seeliger, A. & Mühlhäuser, M. (2018). Analyzing business process anomalies using autoencoders. *Machine Learning*, 107(11): 1875-1893. <https://doi.org/10.1007/s10994-018-5702-8>
- Nolle, T., Luettggen, S., Seeliger, A. & Mühlhäuser, M. (2019). Binet: Multi-perspective business process anomaly classification. *Information Systems*, 103: 101458. <https://doi.org/10.1016/j.is.2019.101458>
- Nolle, T., Seeliger, A. & Mühlhäuser, M. (2016). *Unsupervised anomaly detection in noisy business process event logs using denoising autoencoders*. Discovery Science: 19th International Conference, DS 2016, Bari, Italy, October 19–21, 2016, Proceedings 19.
- Pauwels, S. & Calders, T. (2019). Detecting anomalies in hybrid business process logs. *ACM SIGAPP Applied Computing Review*, 19(2): 18-30. <https://doi.org/10.1145/3357385.3357387>
- Petrak, L. & Lorenz, R. (2020). *Detecting Infrequent Behavior in Event Logs using Statistical Inference*. ATAED@ Petri Nets.
- Ramos-Gutiérrez, B., Varela-Vaca, Á.J., Ortega, F.J., Gómez-López, M.T. & Wynn, M. T. (2021). *A NLP-oriented methodology to enhance event log quality*. Enterprise, Business-Process and Information Systems Modeling: 22nd International Conference, BPMDS 2021, and 26th International Conference, EMMSAD 2021, Held at CAiSE 2021, Melbourne, VIC, Australia, June 28–29, 2021, Proceedings.
- Redman, T.C. (1998). The impact of poor data quality on the typical enterprise. *Communications of the ACM*, 41(2): 79-82. <https://doi.org/10.1145/269012.269025>
- Reinkemeyer, L. (2020). *Process Mining in Action, Principles, Use Cases and Outlook*. Springer. <https://doi.org/10.1007/978-3-030-40172-6>
- Rogge-Solti, A., Mans, R.S., van der Aalst, W.M. & Weske, M. (2013a). *Improving documentation by repairing event logs*. IFIP Working Conference on The Practice of Enterprise Modeling.
- Rogge-Solti, A., Mans, R.S., van der Aalst, W.M. & Weske, M. (2013b). *Repairing event logs using stochastic process models*. On the Move to Meaningful Internet Systems: OTM 2013 Workshops.
- Sadeghianasl, S., ter Hofstede, A.H., Suriadi, S. & Turkay, S. (2020). *Collaborative and interactive detection and repair of activity labels in process event logs*. 2020 2nd International Conference on Process Mining (ICPM).
- Sadeghianasl, S., ter Hofstede, A.H., Wynn, M.T. & Suriadi, S. (2019). *A contextual approach to detecting synonymous and polluted activity labels in process event logs*. On the Move to Meaningful Internet Systems: OTM 2019 Conferences: Confederated International Conferences: CoopIS, ODBASE, C&TC 2019, Rhodes, Greece, October 21–25, 2019, Proceedings.
- Sani, M.F. (2020). *Preprocessing Event Data in Process Mining*. International Conference on Advanced Information Systems Engineering.
- Sani, M.F., Boltenhagen, M. & van der Aalst, W. (2019). Prototype selection based on clustering and conformance metrics for model discovery. *arXiv preprint arXiv: 1912.00736*.

<https://doi.org/10.48550/arXiv.1912.00736>

- Sani, M.F., van Zelst, S.J. & van der Aalst, W.M. (2017). *Improving process discovery results by filtering outliers using conditional behavioural probabilities*. International Conference on Business Process Management.
- Sani, M.F., van Zelst, S.J. & van der Aalst, W.M. (2018a). *Applying sequence mining for outlier detection in process mining*. OTM Confederated International Conferences on the Move to Meaningful Internet Systems.
- Sani, M.F., van Zelst, S.J. & van der Aalst, W.M. (2018b). *Repairing outlier behaviour in event logs*. International Conference on Business Information Systems.
- Sani, M.F., van Zelst, S.J. & van der Aalst, W.M. (2019). Repairing outlier behaviour in event logs using contextual behaviour. *Enterprise Modelling and Information Systems Architectures (EMISAJ)*, 14(5): 1-24.
- Shami Zanjani, M., Nabibi, F. & Irandoust, S. (2020). *Digital Leadership, a Guide to the Transformation of Organizations in the Digital Age*. Tehran: Aryanaghaham Publication. [in persian]
- Sim, S., Bae, H. & Choi, Y. (2019). *Likelihood-based multiple imputation by event chain methodology for repair of imperfect event logs with missing data*. 2019 International Conference on Process Mining (ICPM).
- Song, S., Cao, Y. & Wang, J. (2016). Cleaning timestamps with temporal constraints. *Proceedings of the VLDB Endowment*, 9(10): 708-719. <https://doi.org/10.14778/2977797.2977798>
- Song, S., Gao, Y., Wang, C., Zhu, X., Wang, J. & Philip, S.Y. (2017). Matching heterogeneous events with patterns. *IEEE Transactions on Knowledge and Data Engineering*, 29(8): 1695-1708. <https://doi.org/10.1109/TKDE.2017.2690912>
- Song, W., Jacobsen, H.-A. & Zhang, P. (2019). *Self-Healing Event Logs*. IEEE Transactions on Knowledge and Data Engineering. <https://doi.org/10.1109/TKDE.2019.2956520>
- Song, W., Xia, X., Jacobsen, H.-A., Zhang, P. & Hu, H. (2015). *Heuristic recovery of missing events in process logs*. 2015 IEEE International Conference on Web Services.
- Steeman, W. (2013). *BPI Challenge 2013, incidents*. 4TU. ResearchData. Dataset. <https://doi.org/10.4121/uuid:500573e6-accc-4b0c-9576-aa5468b10cee>
- Sun, X., Hou, W., Yu, D., Wang, J. & Pan, J. (2019). *Filtering out noise logs for process modelling based on event dependency*. 2019 IEEE International Conference on Web Services (ICWS).
- Suriadi, S., Andrews, R., ter Hofstede, A.H. & Wynn, M.T. (2017). Event log imperfection patterns for process mining: Towards a systematic approach to cleaning event logs. *Information systems*, 64: 132-150. <https://doi.org/10.1016/j.is.2016.07.011>
- Tax, N. (2019). Mining insights from weakly-structured event data. *arXiv preprint arXiv:1909.01421*. <https://doi.org/10.48550/arXiv.1909.01421>
- Tax, N., Alasgarov, E., Sidorova, N., Haakma, R. & van der Aalst, W.M. (a2019). Generating time-based label refinements to discover more precise process models. *Journal of Ambient Intelligence and Smart Environments*, 11(2): 165-182. <https://doi.org/10.3233/AIS-190519>
- Tax, N., Sidorova, N. & van der Aalst, W.M. (b2019). Discovering more precise process models from event logs by filtering out chaotic activities. *Journal of Intelligent Information Systems*, 52(1): 107-139. <https://doi.org/10.1007/s10844-018-0507-6>

- van Cruchten, R. (2019). *Data Quality in Process Mining: A Rule-based Approach*. Interactive Process Mining in Healthcare: 53-79. Springer, Cham.
<http://dx.doi.org/10.1007/978-3-030-53993-1-5>
- Van Der Aalst, W. (2011). *Process Mining: Discovery, Conformance and Enhancement of Business Processes*. Springer. <https://doi.org/10.1007/978-3-642-19345-3>
- Van der Aalst, W. (2015). *Process Mining: Discovery, Conformance and Enhancement of Business Process*. Translated by: S.H. Siadat & R. Hemmati Goshtasb. Tehran: Shahid Beheshti Uni. Press. [in persian]
- Van Der Aalst, W. (2016). *Process Mining, Data science in action*. Springer.
<https://doi.org/10.1007/978-3-662-49851-4>
- Van Der Aalst, W., Adriansyah, A., De Medeiros, A.K.A., Arcieri, F., Baier, T., Blickle, T., Bose, J.C., Van Den Brand, P., Brandtjen, R. & Buijs, J. (2011). *Process mining manifesto*. International conference on business process management.
- Van der Aalst, W., Weijters, T. & 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.2004.47>
- van der Aalst, W.M., Bolt, A. & van Zelst, S.J. (2017). RapidProM: mine your processes and not just your data. *arXiv preprint arXiv: 1703.03740*. <https://doi.org/10.48550/arXiv.1703.03740>
- van Dongen, B. (2011). *Hospital log (BPIC 2011)*. 4TU.ResearchData. Dataset.
<https://doi.org/10.4121/uuid:d9769f3d-0ab0-4fb8-803b-0d1120ffcf54>
- Van Dongen, B. (2012). *BPI Challenge 2012, Event log of a loan application process*. 4TU. ResearchData. Dataset.
<https://doi.org/10.4121/uuid:3926db30-f712-4394-aebc-75976070e91f>
- van Dongen, B. (2015). *BPI Challenge 2015*. 4TU. ResearchData. Dataset.
<https://doi.org/10.4121/uuid:31a308ef-c844-48da-948c-305d167a0ec1>
- van Dongen, B. (2017). *BPI Challenge 2017*. 4TU. ResearchData. Dataset.
<https://doi.org/10.4121/uuid:5f3067df-f10b-45da-b98b-86ae4c7a310b>
- van Dongen, B. (2018). *BPI Challenge 2018*. 4TU. ResearchData. Dataset.
https://data.4tu.nl/articles/dataset/BPI_Challenge_2018/12688355
- van Dongen, B. (2019). *BPI Challenge 2019*. 4TU. ResearchData. Dataset.
<https://doi.org/10.4121/uuid:d06aff4b-79f0-45e6-8ec8-e19730c248f1>
- van Scheepstal, S. (2016). *Data quality within process mining in the auditing context* Tilburg University. Netherlands.
- van Zelst, S.J., Sani, M.F., Ostovar, A., Conforti, R. & La Rosa, M. (2018). *Filtering spurious events from event streams of business processes*. International Conference on Advanced Information Systems Engineering.
- van Zelst, S.J., Sani, M.F., Ostovar, A., Conforti, R. & La Rosa, M. (2020). Detection and removal of infrequent behavior from event streams of business processes. *Information systems*, 90: 101451. <https://doi.org/10.1016/j.is.2019.101451>
- Vanbrabant, L., Martin, N., Ramaekers, K. & Braekers, K. (2019). Quality of input data in emergency department simulations: framework and assessment techniques. *Simulation Modelling Practice and Theory*, 91: 83-101. <https://doi.org/10.1016/j.simpat.2018.12.002>

- Verbeek, H., Buijs, J., Van Dongen, B. & van der Aalst, W.M. (2010). Prom 6: The process mining toolkit. *Proc. of BPM Demonstration Track*, 615: 34-39.
- Verhulst, R. (2016). *Evaluating quality of event data within event logs: an extensible framework*. Master's thesis, Eindhoven University of Technology].
- Vidgof, M., Djurica, D., Bala, S. & Mendling, J. (2020). *Cherry-picking from spaghetti: Multi-range filtering of event logs*. Enterprise, Business-Process and Information Systems Modeling: 21st International Conference, BPMDS 2020, 25th International Conference, EMMSAD 2020, Held at CAiSE 2020, Grenoble, France, June 8–9, 2020, Proceedings 21.
- Vijayakamal, M. & Vasumathi, D. (2019). Unsupervised Learning Methods for Anomaly Detection and Log Quality Improvement Using Process Event Log. *International Journal of Advanced Science and Technology*, 29: 1109-1125.
- Walicki, M. & Ferreira, D.R. (2011). Sequence partitioning for process mining with unlabeled event logs. *Data & Knowledge Engineering*, 70(10): 821-841.
<https://doi.org/10.1016/j.datak.2011.05.003>
- Wang, J., Song, S., Lin, X., Zhu, X. & Pei, J. (2015). *Cleaning structured event logs: A graph repair approach*. 2015 IEEE 31st International Conference on Data Engineering.
- Wang, J., Song, S., Zhu, X. & Lin, X. (2013). Efficient recovery of missing events. *Proceedings of the VLDB Endowment*, 6(10): 841-852. <https://doi.org/10.14778/2536206.2536212>
- Wang, J., Song, S., Zhu, X., Lin, X. & Sun, J. (2016). Efficient recovery of missing events. *IEEE Transactions on Knowledge and Data Engineering*, 28(11): 2943-2957.
- Wang, L.-l., Fang, X.-W., Asare, E. & Huan, F. (2021). *An Optimization Approach for Mining of Process Models with Infrequent Behaviors Integrating Data Flow and Control Flow*. Scientific Programming, 2021. <https://doi.org/10.1155/2021/8874316>
- Wynn, M.T. & Sadiq, S. (2019). *Responsible process mining-a data quality perspective*. Business Process Management: 17th International Conference, BPM 2019, Vienna, Austria, September 1–6, 2019, Proceedings 17.
- Xia, X., Song, W., Chen, F., Li, X. & Zhang, P. (2016). *Effa: A ProM plugin for recovering event logs*. Proceedings of the 8th Asia-Pacific Symposium on Internetware.
- Xu, J. & Liu, J. (2019). A profile clustering-based event logs repairing approach for process mining. *IEEE Access*, 7: 17872-17881. <https://doi.org/10.1109/ACCESS.2019.2894905>
- Yi, G. & Peng, Z. (2019). Novel Approach to Discover Precise Process Model by Filtering out Log Chaotic Activities. *Journal of Computers*, 30(4): 140-150.
- Zerbino, P., Stefanini, A. & Aloini, D. (2021). Process science in action: A literature review on process mining in business management. *Technological Forecasting and Social Change*, 172: 121021. <https://doi.org/10.1016/j.techfore.2021.121021>
- Zhu, X., Song, S., Lian, X., Wang, J. & Zou, L. (2014). *Matching heterogeneous event data*. Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data.
- Zhu, X., Song, S., Wang, J., Philip, S.Y. & Sun, J. (2014). *Matching heterogeneous events with patterns*. 2014 IEEE 30th International Conference on Data Engineering.

پیوست الف

جدول ۱- پژوهش‌های بررسی شده مربوط به گروه چارچوب‌های ارزیابی کیفیت داده

هدف	جنبه‌های کیفیت داده مورد تاکید	مجموعه داده استاندارد مورد استفاده	ابزار پیاده‌سازی	منبع
ارائه تعدادی اصول راهنما برای تمایز سطح کیفیت گزارش‌های رویداد در ۵ سطح متمایز از عالی تا ضعیف	C1, Ti1	-	-	Van Der Aalst (& et al., 2011)
ارائه یک روش ساختارمند به جهت شناسایی مشکلات کیفیت داده در ۱۰ دسته‌بندی مختلف به همراه تشریح تاثیر هر کدام از آن‌ها بر روی فرایندکاوی	C1, C2, C3, E1, E4, E5, Ti1, Ti2, Ti3	-	-	Bose & et al., (2013)
ارائه یک دسته‌بندی جدید از ابعاد کیفیت داده در فرایندکاوی، در زمینه حسابداری و ارزیابی آن‌ها توسط خبرگان، با استفاده از انجام مصاحبه. نویسندگان نتیجه گرفته‌اند که همه ابعاد فرایندکاوی در زمینه حسابداری تاثیرگذار بوده و می‌تواند یا ناشی از انحراف در فرایندهای کسب‌وکار شرکت‌ها باشد، یا ناشی از مشکلات مربوط به ثبت رویدادها در گزارش رویداد	C1, C2, E5	-	-	van Scheepstal, (2016)
در این مقاله با استفاده از رویکردهای ارزیابی کیفیت داده موجود در ادبیات، یک چارچوب قابل توسعه که ترکیبی از بهترین جنبه‌های رویکردهای گذشته بوده و نیز تعریف روشنی از هر بُعد کیفیت داده و معیار ارزیابی متناظر با آن ارائه شده است.	E2, E3, E4	BPIC 2011, BPIC 2012, BPIC 2015	ProM	(Verhulst, 2016)
معرفی یک مدل ارزیابی کیفیت داده چند سطحی با قابلیت یکپارچه‌سازی ابعاد مختلف، به منظور ایجاد درک کلی از کیفیت گزارش رویداد قبل از هر تحلیل، و پروژه فرایندکاوی به همراه تعریف معیارهای ارزیابی لازم	E1, E2, E3, E4	BPIC 2012	ProM	Kherbouche & (et al., 2016)
در این مقاله نویسندگان در مورد یک چارچوب مفهومی صحبت می‌کنند تا موارد مختلف مربوط به کیفیت داده را مشخص کرده و به یکدیگر پیوند دهند، تا درک کلی نسبت به کیفیت داده‌های ورودی به دست آید.	E1, E2, Ti1	-	-	Lu & Fahland, (2017)
معرفی و شناسایی الگوهای رایج مشکلات کیفیت داده در ۱۱ دسته‌بندی متمایز، به همراه ارائه ویژگی‌ها و تاثیر آن‌ها بر روی نتایج فرایندکاوی. برای هر الگو یک مثال از دنیای واقعی و راه‌حل مقابله با آن شرح داده شده است.	A2, A3, C1, C2, C3, E1, E2, E3, E4, E5	-	-	Suriadi & (et al., 2017)
نویسندگان یک زبان پرس‌وجو تحت عنوان «پرس‌وجوی گزارش رویداد برای نقص‌ها (QUELI)» ارائه داده‌اند که وظیفه آن شناسایی مشکلات کیفیت موجود در یک گزارش رویداد براساس ۱۱ الگوی معرفی شده در مطالعه (Suriadi & et al., 2017) می‌باشد که می‌تواند ۵ مورد از الگوهای معرفی شده در مطالعه قبلی را در یک لاگ شناسایی نماید.	A3, Ti1, Ti2	-	-	Andrews & (et al., 2018)

هدف	جنبه‌های کیفیت داده مورد تاکید	مجموعه داده استاندارد مورد استفاده	ابزار پیاده‌سازی	منبع
توسعه یک چارچوب کیفیت داده به نام «CP-DQF» به منظور شناسایی، مدیریت و کاهش مشکلات کیفیت داده در گزارش‌های رویداد مربوط در زمینه سلامت	T1, T2	-	Microsoft SQL Server	Fox & et al., (2018)
ارائه رویکردی مفهومی برای کاهش مشکلات کیفیت داده در داده‌های مراقبت سلامت از طریق یک سامانه مکان‌یابی داخلی ^۱ (ILS)، که براساس داده‌هایی که ILS فراهم می‌کند، مشکلات داده‌ای جدیدی را کشف می‌نماید. داده‌های ILS می‌توانند نقشی مهم در کاهش مشکلات کیفیت داده ایفا کنند که در سامانه‌های مراقبت سلامت مهم هستند.	A1, A2, A3, C1, C2, C3, E4, E5, Ti1, Ti2, Ti3, Ti4	-	-	(Martin, 2018)
ارائه یک رویکرد جدید برای چگونگی ایجاد یک تحلیل فرایندکاوی مناسب براساس شناسایی دقیق مشکلات مربوط به کیفیت داده، قبل از اعمال هرگونه روش فرایندکاوی به جهت کاهش هزینه و اتلاف زمان	E2, E3, E4, E5	-	-	Andrews & (et al., 2019)
هدف این مقاله ایجاد روشی نظام‌مند، خودکار و شفاف برای شناسایی و حل مشکلات کیفیت گزارش‌های رویداد در سازمان‌ها می‌باشد. رویکرد فوق به گونه‌ای طراحی شده است که قابلیت استفاده عمومی داشته باشد.	C2, E1, E5	-	-	van Cruchten, (2019)
ارائه یک چارچوب کیفیت داده برای دسته‌بندی مشکلات کیفیت داده در داده‌های رویداد مربوط به پرونده‌های الکترونیکی مراقبت سلامت، ^۲ به منظور انجام بهتر تحلیل‌های شبیه‌سازی. ^۳ هدف این پژوهش شناسایی و کمی‌سازی مشکلات کیفیت داده شناسایی شده در زمینه سلامت است.	E2, E4, E5	-	-	Vanbrabant & (et al., 2019)
ارائه یک رویکرد متأثر از متدولوژی I* [*] برای ارزیابی مشکلات کیفیت داده در داده‌های مراقبت سلامت	E1, Ti1, Ti2	-	-	Kurniati & (et al., 2019)
ارائه یک رویکرد تعاملی برای ارزیابی کیفیت داده‌ها، هم در سطح گزارش رویداد و هم در سطح کشف فرایند، برای شناسایی ناسازگاری‌های پنهان موجود در لاگ، به همراه یک روش اکتشافی برای پاکسازی داده‌ها	Ti1, Ti2, Ti3, Ti4	-	-	(Martin, 2019)
در این مقاله به نوع جدیدی از فرایندکاوی با عنوان فرایندکاوی پاسخگو اشاره شده است. نتایج نشان داد که برای حرکت به سمت فرایندکاوی پاسخگو، کسب آگاهی از کیفیت داده‌ها و کاهش نتایج دارای اشتباه، بسیار حائز اهمیت است. در این پژوهش بر روی گزارش‌های رویداد به عنوان ورودی اصلی روش‌های فرایندکاوی تمرکز شده است.	E1, E2, E3, E4, E5	-	-	Wynn & Sadiq, (2019)

http://stim.gom.ac.ir

1. Insert location system
2. Electronic healthcare records
3. Simulation

منبع	ابزار پیاپی‌سازی	مجموعه داده استاندارد مورد استفاده	جنبه‌های کیفیت داده مورد تاکید	هدف
Emamjome & (et al., 2020)	-	-	A6, E4, E5	معرفی چارچوبی به نام Odigos برای شناخت و ارزیابی مشکلات کیفیت داده در فرایندکاوی از دیدگاهی جدید که می‌توان از آن هم به صورت پیش‌آگهی (پیش‌بینی مشکلات احتمالی کیفیت) و هم تشخیصی (شناسایی دلایل ریشه‌ای مشکلات مربوط به کیفیت) استفاده نمود. در این پژوهش کیفیت داده در فرایندکاوی از چهار دیدگاه بررسی می‌شود: (۱) تعاملات بین دنیای اجتماعی و شخصی، (۲) تعاملات بین دنیای فردی و تکنولوژی (نرم‌افزار و سخت‌افزار) (۳) تعامل بین دنیای اجتماعی و تکنولوژی این چارچوب با الهام از پژوهش (Mingers & Willcocks, 2014) توسعه داده شده است.
Andrews,) Emamjome, & (et al., 2020	-	-	C2, C3, E3, E4	معرفی چارچوبی مفهومی برای ارزیابی کیفیت گزارش‌های رویداد، در پروژه‌های فرایندکاوی. نویسندگان چارچوب فوق را از طریق مصاحبه با متخصصان دانش فرایندکاوی مورد ارزیابی قرار داده‌اند. سه بخش اصلی این چارچوب شامل: موارد اجتماعی (موقعیت سازمانی)، موارد فردی (مشارکت در فرایند، نگهداری داده‌ها) و سامانه‌های نرم‌افزاری و سخت‌افزاری بوده که در میان این سه ضلع فرایندها، داده‌های رویداد و گزارش‌های رویداد قرار دارند که نشان‌دهنده این است که کیفیت داده یک موضوع چند بُعدی است. هدف این چارچوب ایجاد راهی برای شناسایی و پیش‌بینی موارد مربوط به کیفیت داده گزارش‌های رویداد در یک سازمان است. نسخه کامل چارچوب پیشنهادی در پژوهش (Emamjome & et al., 2020) ذکر شده است.
Fischer & (et al., 2020)	ProM	-	A1, A2, A3, C1, C2, C3, E4, E5, Ti1, Ti2, Ti3, Ti4, Ti5	هدف این مقاله بررسی و ارزیابی خودکار مشکلات مربوط به برچسب‌های زمانی، به همراه ۱۵ معیار جهت کمی‌سازی آن‌ها است.
Cappiello & (et al., 2022)	-	-	T2, Ti1	در این مقاله رویکردی جدید برای قابل اندازه‌گیری کردن شاخص‌های عملکرد فرایند ^۱ (PPI) برای ارزیابی کارایی و اثربخشی یک فرایند کسب‌وکار ارائه شده است. نویسندگان بیان می‌کنند که کیفیت داده‌های رویداد ذخیره شده، نقش مهمی برای اندازه‌گیری PPI دارند. به همین جهت آن‌ها یک چارچوب جدید برای پیوند بین معیارهای PPI و کیفیت گزارش‌های رویداد ارائه می‌دهند.

http://stlm.gom.ac.ir

جدول ۲- پژوهش‌های بررسی شده مربوط به دسته پیش‌پردازش

منبع	ابزار پیاده‌سازی	مجموعه داده استاندارد مورد استفاده	جنبه‌های کیفیت داده مورد تاکید	هدف
Ingvaldsen & (Gulla, 2007)	ProM	-	E3	ارائه یک روش پیش‌پردازش برای جمع‌آوری داده‌ها از جداول پراکنده اطلاعاتی در یک سامانه ERP مانند SAP
Ferreira & (Gillblad, 2009)	-	-	C1	معرفی یک رویکرد احتمالی برای کشف مدل فرایند از گزارش‌های رویدادی که فاقد برچسب هستند.
Walicki & (Ferreira, 2011)	-	-	C1	پیاده‌سازی یک روش کارا برای شناسایی شناسه مورد‌های فاقد برچسب
Ly & et al., (2012)	ProM	-	T1, T2, C2	ارائه مفهومی به نام «پالایش معنایی لاگ» ^۱ به عنوان یک مرحله پیش‌پردازش برای پاک‌سازی لاگ براساس محدودیت‌های ^۲ مربوط به زمینه کسب‌وکار. این فعالیت برای افزایش دقت مدل فرایند استخراج شده از لاگ انجام می‌پذیرد.
Leemans & (et al., 2013)	ProM	BPIC 2011, BPIC 2012	T1	کشف یک مدل فرایند درست و کامل، ^۳ با وجود رفتارهای کم‌تکرار
Mueller- Wickop & (Schultz, 2013)	ProM	-	E3	ارائه یک روش پیش‌پردازش برای مرتب نمودن رویدادهای موجود در یک سامانه ERP، با استفاده از تحلیل دانش نهفته در فعالیت‌های مربوط به حسابداری
(Baier, 2015)	ProM	-	E3, A4	همپوشانی رویدادهای کسب‌وکار ذخیره شده در یک گزارش رویداد با مدل فرایند پایه ^۴ داده شده، براساس استفاده از بررسی ویژگی‌های رفتاری ^۵ و زبانشناختی ^۶ مدل‌های فرایند دانش موجود در مستندات سازمان
Cheng & (Kumar, 2015)	-	-	E4	این مقاله روشی کامل‌تر نسبت به پژوهش گذشته نویسندگان (Bayomie & et al., 2016) ارائه می‌دهد. در این مطالعه روشی برای حذف داده‌های آشفته از لاگ، براساس قوانین استخراج شده از لاگ ورودی ارائه شده است.
Bayomie & (et al., 2016)	-	-	E2	معرفی یک روش سه مرحله‌ای به منظور یافتن شناسه نمونه فرایندها ^۷ برای رویدادهای فاقد برچسب در فرایندهای کسب‌وکار چرخه‌ای ^۸

1. Semantic log purging
2. Constraints
3. Sound
4. Reference model
5. Behavioral
6. Linguistic
7. Case id
8. Cyclic

منبع	ابزار پیاده‌سازی	مجموعه داده استاندارد مورد استفاده	جنبه‌های کیفیت داده مورد تاکید	هدف
Conforti & (et al., 2016)	ProM	BPIC 2012	E4	هدف این مقاله ارائه یک روش خودکار به نام AFA برای ایجاد مدل‌های فرایندی با کیفیت بالا، با وجود داده‌های نویزی می‌باشد. این عمل از طریق فیلتر کردن رفتارهای نادرست و نامکرر از گزارش‌های رویداد صورت می‌گیرد.
(Lu, 2016)	ProM	—	E3, A4	حل مشکل برجسب‌های نامناسب (در اینجا تکراری) با استفاده از یک روش پیش‌پردازش، به منظور شناسایی و پالایش برجسب‌های نامناسب
Lu, Fahland & van der Aalst, b (2016)	ProM	Road Fines	E4, Ti1	معرفی یک پلاگین به نام Log to Model Explorer که قابل استفاده در چارچوب برنامه ProM باشد، با هدف خوشه‌بندی دنباله‌ها، حذف رفتارهای کم‌تکرار و اصلاح برجسب‌های زمانی
Ayo & et al.,) (2017)	ProM	—	T2	بررسی وابستگی بین فعالیت‌ها و ارائه یک روش احتمالی برای پیش‌بینی فعالیت‌های گمشده در دنباله‌های ناکامل، با استفاده از توابع امتیازدهی بیزی ^۱
Mannhardt,) (2017)	ProM	Road Fines	E4	معرفی روشی به نام DHM برای کشف جریان کنترل یک فرایند و تشخیص مسیرهای آشفته از مسیرهای کم‌تکرار، با استفاده از تحلیل وابستگی بین رویدادها
Sani & et al.,) (2017)	ProM RapidProM	BPIC 2012, BPIC 2017	E4	توسعه یک روش پالایش بر پایه احتمالات شرطی، ^۲ برای تعیین احتمال وقوع یک فعالیت براساس رفتار فعالیت‌های نزدیک به آن. اگر احتمال یک فعالیت کمتر از مقدار آستانه باشد، آن فعالیت رفتار پرت در نظر گرفته می‌شود.
Song & et al.,) (2017)	—	—	E3	نسخه اولیه این مقاله در سال ۲۰۱۴ ارائه شده است (Zhu, Song, Wang & et al., 2014) و این مطالعه توسعه‌یافته مقاله قبلی است. در این مقاله مسأله تطابق از نظر ریاضی فرموله شده و سپس یک چارچوب برای مطابقت رویدادها براساس ساختار رویدادها، معرفی می‌شود.
Gao & et al.,) (2018)	—	—	E2, Ti1, Ti3	توسعه مقاله (Zhu, Song, Lian & et al., 2014) با هدف ایجاد یک روش مطابقت با محاسبه شباهت بین رویدادها و مشخص کردن رویدادهای متناظر
Huang & (et al., 2018)	—	—	A1	روشی برای حذف فعالیت‌های پرت براساس تکنیک آماری IQR ^۳

http://stlm.gom.ac.ir

1. Bayesian Scoring Functions
2. Conditional probabilities
3. Interquartile range

منبع	ابزار پیاده‌سازی	مجموعه داده استاندارد مورد استفاده	جنبه‌های کیفیت داده مورد تاکید	هدف
Sani & et al., (2018a)	-	BPIC 2012, BPIC 2017, BPIC 2018, Hospital Billing, Sepsis	E4	در این پژوهش روشی سه مرحله‌ای برای استخراج قوانین و الگوهای دنباله‌ها از لاگ ارائه شده، سپس شناسایی نقاط پرت برای دنباله و در مرحله بعد دنباله‌های با رفتارهای پرت از لاگ حذف می‌شوند.
Chapela-) Campa & (et al., 2019	ProM RapidProM	BPIC 2011, BPIC 2012, BPIC 2013, BPIC 2105, Sepsis	E4	ساده‌سازی لاگ توسط الگوریتمی به نام WoSimp. برای حفظ مقادیر اصلی و حذف مقادیر کم‌تکرار که منجر به ساده‌سازی لاگ و کاهش پیچیدگی مدل فرایند استخراج می‌شود.
Chen & et al., (2019)	-	-	A2, A3, A4, Ti3, Ti5	ارائه یک چارچوب پیش‌پردازش، برای حل مشکلات کیفیت داده گزارش‌های رویداد، برای تحلیل اطلاعات پروژه‌های ساختمانی و عمرانی.
Sun & et al., (2019)	ProM	-	E4	تعریف وابستگی بین رویدادها، به صورتی که هرچه وابستگی بین رویدادها بیشتر باشد، آن‌ها عادی‌تر بوده و هرچه وابستگی بین آن‌ها کمتر باشد، در دسته رویدادهای آشفته قرار می‌گیرند.
(Tax, 2019)	-	-	A5	این پایان‌نامه به کشف فرایند بر مبنای گزارش‌های رویداد دارای مشکل کیفیت داده (ساختار ضعیف) می‌پردازد، که در آن چند روش برای پیش‌پردازش گزارش رویداد ارائه می‌شود.
Tax,) Alasgarov & (et al., a 2019	ProM	-	Ti1, Ti2	ارائه یک روش خودکار برای اصلاح برچسب‌های یک گزارش رویداد، به منظور کشف مدل‌های فرایند دقیق‌تر. نویسندگان چهار استراتژی موثر در این زمینه را براساس داده‌های واقعی خانه‌های هوشمند، مورد بحث و بررسی قرار داده‌اند.
Tax, Sidorova) (& et al., b2019	RapidProM	-	A5	ارائه و بررسی چهار روش پیشنهادی برای حذف فعالیت‌های آشفته، با در نظر گرفتن میزان رخداد آن‌ها
Sani,) Boltenhagen & (et al., 2019	ProM RapidProM	BPIC 2012, BPIC 2018, BPIC 2019, Hospital Billing, Road Fines, Sepsis	E1	ارائه یک روش پیش‌پردازش چهار مرحله‌ای شامل خوشه‌بندی دنباله‌های موجود در گزارش رویداد، برای ساخت نمونه‌های اولیه ورودی، کشف مدل براساس نمونه‌های ساخته شده، ارزیابی کیفیت نتایج خروجی و در انتها، تکرار سه روند قبلی تا حصول کسب بهترین نتیجه
Yi & Peng, (2019)	ProM	-	E4	ارائه یک روش برای تحلیل وابستگی دنباله‌ها، براساس کشف رابطه‌های سببی دو طرفه ^۱ میان فعالیت‌ها، از طریق بررسی مدل فرایند و گزارش رویداد

منبع	ابزار پیاده‌سازی	مجموعه داده استاندارد مورد استفاده	جنبه‌های کیفیت داده مورد تاکید	هدف
Bernard & Andritsos, (2020)	Python	BPIC 2012, BPIC 2013, BPIC 2015, BPIC 2017, BPIC 2018, BPIC 2019	T2	معرفی مفهومی جدید به نام Truncated Trace که در واقع همان دنباله‌های ناکامل هستند. در این مطالعه یک طبقه‌بندی برای فیلتر کردن اینگونه دنباله‌ها و در نتیجه افزایش دقت مدل‌های فرایند توسعه داده شده است. برای این منظور از XGBoost و شبکه‌های عصبی استفاده شده است.
Andrews, van Dun & et al., (2020)	MySQL SQL Server	-	A2, E5, Ti2, Ti4	معرفی رویکردی به نام RDB2Log برای استخراج داده‌های رویداد، با استفاده از محدودیت‌های مربوط به داده‌های ذخیره شده در پایگاه داده رابطه‌ای ^۱ ، به صورت نیمه‌خودکار
Díaz- Rodriguez & Hernández, (2020)	-	-	T1, T2, Ti5	ارائه روشی برای ایجاد یک گزارش رویداد با کیفیت از داده‌های رویداد ذخیره شده در سامانه‌های اطلاعاتی براساس نیاز و اهداف کاربر
Huang & (et al., 2020)	-	Hospital billing, Sepsis	E4	ارائه روشی برای حذف رویدادهای آشفته / کم‌تکرار از گزارش رویداد، به واسطه محاسبه حداقل انتظارات ^۲ ، به عنوان یک مقدار آستانه برای کشف رفتارهای کم‌تکرار به منظور بهبود روش‌های کشف فرایند
Petrak & (Lorenz, 2020)	Java	BPIC 2012	E4	حذف داده‌های آشفته که در این پژوهش از آن‌ها تحت عنوان حذف وابستگی فعالیت‌های موجود در دنباله‌های کم‌تکرار یاد می‌شود. رویکرد معرفی شده براساس روش‌های آماری مبتنی بر آزمون فرضیه است که منجر به ایجاد یک‌سری کران برای طبقه‌بندی رفتارهای نامتداول خواهد شد.
(Sani, 2020)	ProM	BPIC 2012, BPIC 2018, BPIC 2019, Hospital Billing, Road Fines, Sepsis	E1	اعمال پیش‌پردازش روی لاگ ورودی برای کاهش اندازه و پیچیدگی آن. این فعالیت با استفاده از تجمیع رویدادهای مشابه، از طریق تجمیع دنباله‌ها، فعالیت‌ها و یا هر دو صورت می‌پذیرد.
van Zelst & (et al., 2020)	-	Sepsis	E4	این مطالعه توسعه مقاله قبلی نویسندگان است (van Zelst et al., 2018). ارائه یک دسته‌بندی جدید برای طبقه‌بندی رفتارها در جریان کنترل، برای شناسایی رفتارهای پرت

http://stlm.gom.ac.ir

1. Relational database
2. Minimum expectations

منبع	ابزار پیاده‌سازی	مجموعه داده استاندارد مورد استفاده	جنبه‌های کیفیت داده مورد تاکید	هدف
Vidgof & (et al., 2020)	PM4Py	Sepsis	E4	ارائه یک روش پالایش چند محدوده‌ای ^۱ که مقادیر آن توسط کاربر مشخص شده است، که براساس تعداد رخداد رویدادها بر مبنای یک بازه تعیین می‌شود.
Chen & et al.,) (2021	-	-	A2	در این مقاله با استفاده از سه رویکرد جریان کنترل، ویژگی‌های عددی و NLP، برچسب فعالیت‌های مشابه شناسایی می‌شوند و سپس با اجماع نتایج حاصل، از سه رویکرد فوق و ارزیابی آن در مورد برچسب فعالیت‌های اضافی (تکراری) تصمیم‌گیری می‌شود. عمل تصمیم‌گیری توسط تحلیل‌گر صورت می‌پذیرد.
Lu & et al.,) (2021	PM4Py	BPIC 2020	E4	ارائه یک روش جدید برای تمایز بین رویدادهای آشفته از رویدادهای کم-تکرار مهم و موثر در فرایند کسب‌وکار، که بر مبنای یک مفهوم احتمالی به نام مسیرهای حداکثر احتمال ^۲ ، با هدف بررسی حالت‌های مختلف فعالیت‌ها در یک فرایند و همچنین مشخص کردن حالت‌های گذار از یک فعالیت به فعالیت دیگر (تمایز بین فعالیت‌های منظم و نامنظم)، رفتارهای موجود در یک دنباله را به دو قسمت سودمند یا غیرمفید تقسیم می‌نماید.
Ramos-) Gutiérrez & (et al., 2021	Java	-	A1	در این مقاله به داده‌های ذخیره شده در یک گزارش رویداد، به عنوان یک سری کلمات متنی نگاه می‌شود. هدف، بهبود کیفیت گزارش رویداد با استفاده از تکنیک‌های پردازش متن طبیعی یا NLP است، تا با پالایش برچسب‌های گزارش رویداد (مانند برچسب‌های زمانی، نام فعالیت یا شناسه نمونه فرایند) بتوان مدل‌های فرایندی دقیق‌تر و خلاصه‌تری به دست آورد. برای انتخاب تکنیک NLP متناسب با این هدف، باید سطح کیفیت گزارش رویداد مشخص شده و انتظارات مدنظر، مورد ارزیابی قرار گیرد که بر این اساس، یک سامانه پشتیبان تصمیم، برای پیشنهاد بهترین تکنیک NLP هم توسعه داده شده است. بدین منظور نویسندگان متدولوژی‌ای به نام LOADING-NLP معرفی نموده‌اند که به تحلیل‌گران برای تصمیم‌گیری بهتر کمک می‌نماید. برای ارزیابی کیفیت گزارش رویداد در روش فوق، معیارهای اندازه‌گیری مناسب نیز پیشنهاد شده است.

<http://stlm.gom.ac.ir>

1. Multi-range Filtering
2. Maximum probability path

منبع	ابزار پیاده‌سازی	مجموعه داده استاندارد مورد استفاده	جنبه‌های کیفیت داده مورد تأکید	هدف
Wang & et al.,) (2021	-	-	E4	در این مقاله آمده است که شناسایی و حذف رویدادهای کم‌تکرار براساس محاسبه تعداد رخداد صحیح نیست، چون همه این رویدادها نادرست نیستند و به علت شرایط یک فرایند کسب‌وکار ایجاد شده‌اند. بدین ترتیب برای ایجاد تمایز بین رفتارهای کم‌تکرار صحیح و نادرست، یک روش جدید بر مبنای الگوهای مکرر و جریان داده ایجاد می‌شود، تا رویدادهای با بسامد پایین را براساس تحلیل الگوی رفتاری و همچنین تأثیر آن بر جریان داده‌های یک فرایند، استخراج نمود.

جدول ۳- پژوهش‌های بررسی شده مربوط به گروه کشف ناهنجاری

منبع	ابزار پیاده‌سازی	مجموعه داده استاندارد مورد استفاده	جنبه‌های کیفیت داده مورد تأکید	هدف
Ghionna &) (et al., 2008	ProM	-	E4	ارائه روشی برای کشف ناهنجاری از طریق تمایز بین رویدادها، با رفتارهای مکرر که منجر به کشف یک مسیر فرایند ناهنجار می‌شود. برای این منظور از یک حد آستانه برای تعیین حداقل فرکانس رخداد فعالیت‌ها استفاده می‌شود.
Bezerra & et al.,) (2009	ProM	-	E4	تعیین وابستگی بین فعالیت‌ها، با توجه به زمینه کسب‌وکار و کشف و فیلتر کردن نمونه فرایندها و فعالیت‌های غیرمرتبط
Bezerra &) (Wainer, 2013	-	-	E4	معرفی و بررسی چهار رویکرد کشف ناهنجاری
Böhmer &) Rinderle-Ma, (2016	-	-	E4	توسعه یک روش احتمالی بر پایه چشم‌اندازهای متنوع فرایندکاوی (جریان کنترل، دیدگاه سازمانی یا منبع) برای فهم اینکه رویدادی توسط رویدادی دیگر دنبال می‌شود، یا خیر؛ سپس تشخیص و حذف رویدادهای ناهنجار
Lu, Fahland, van) den Biggelaar &) (et al., a2016	ProM	BPIC 2015	E4	ارائه یک رویکرد کشف انحرافات مبتنی بر گزارش رویداد، بدون استفاده از مدل فرایند مرجع
Nolle & et al.,) (2016	-	-	E4	معرفی یک روش بر مبنای شبکه‌های عصبی خودرمزگذار ^۱ برای استخراج وابستگی‌ها و ساختارهای نهان داخل لاگ، جهت کشف رفتارهای ناهنجار
Alharbi & et al.,) (2017	ProM	-	E4	توسعه روشی برای شناسایی الگوهای رفتاری مکرر و حذف رویدادهای پرت، با استفاده از تکنیک‌های فاصله، برای تعیین آستانه انتخاب رویداد پرت

<http://stlm.gom.ac.ir>

هدف	جنبه‌های کیفیت داده مورد تاکید	مجموعه داده استاندارد مورد استفاده	ابزار پیاده‌سازی	منبع
ارائه ابزاری به نام Log Pattern که بر مبنای روشی شبه-نظارتی ^۱ برای شناخت الگوهای پیچیده در یک گزارش رویداد، توسعه داده شده است. با استفاده از روش پیشنهادی، برخی از مشکلات کیفیت داده در یک گزارش رویداد، قابل تشخیص است.	A3	-	-	Lu & et al.,) (2017
این مطالعه توسعه یافته روش معرفی شده در مقاله (Nolle & et al., 2016) است. در روش معرفی شده جدید، پالایش‌های بیشتر برای افزایش کارایی در نظر گرفته شده و روش فوق، روی مجموعه داده‌های بیشتری آزمایش شده است.	E4	BPIC 2012, BPIC 2013, BPIC 2015, BPIC 2017	-	Nolle & et al.,) (2018
شناسایی و حذف رویدادهای اشتباه ^۲ از جریان رویدادها ^۳	E4	Sepsis	ProM RapidProM	van Zelst &) (et al., 2018
در این مقاله یک شبکه عصبی به نام BINet، برای کشف ناهنجاری‌های چندمتغیره به صورت همزمان در گزارش‌های رویداد فرایندهای کسب و کار معرفی شده است. نویسندگان در مقاله (Nolle et al., 2019) یک رویکرد جامع‌تر بر مبنای روش BINet ارائه کرده‌اند.	E4, C2	-	-	Nolle & et al.,) (2018
کشف ناهنجاری‌ها بدون استفاده از دانش پایه و یادگیری براساس داده‌های موجود در لاگ، با استفاده از شبکه‌های بیزی پویا ^۴	E4, C2	-	-	Pauwels &) (Calders, 2019
در این مقاله بر روی دو الگوی نقص برچسب فعالیت‌ها، یعنی برچسب‌های مترادف و نادرست تمرکز شده است. نویسندگان با معرفی و استفاده از یک چارچوب به نام «چارچوب زمینه فعالیت» ^۵ برای تعیین زمینه فعالیت‌ها ^۶ و محاسبه معیارهای فاصله، فعالیت‌های مترادف از نظر معنایی و همچنین آلوده ^۷ را مشخص می‌کنند.	A1, A3	Hospital Billing, Sepsis	ProM	Sadeghianasl &) (et al., 2019

<http://stim.gom.ac.ir>

1. Semi-supervised
2. Spurious
3. Event streams
4. Dynamic Bayesian Networks
5. Activity Context Framework
6. Activity context
7. Polluted labels

هدف	جنبه‌های کیفیت داده مورد تاکید	مجموعه داده استاندارد مورد استفاده	ابزار پیاده‌سازی	منبع
ارائه دو الگوریتم بدون نظارت قطعی (SAE-AD) ^۱ و احتمالی (PAE-AD) ^۲ براساس شبکه‌های عصبی مصنوعی (ANN)، برای کشف ناهنجاری در یک گزارش رویداد و حذف آن‌ها از گزارش رویداد و در نتیجه بهبود نتایج فرایندکاوی	E4, T11	BPIC 2015, BPIC 2017, BPIC 2018	-	Vijayakamal & (Vasumathi, 2019)
ارائه رویکردی برای کشف ناهنجاری‌های موجود در یک گزارش رویداد، بدون نیاز به دانش دامنه و مدل مرجع، با استفاده از تکنیک word2vec (یکی از روش‌های NLP)، برای ایجاد یک طبقه‌بندی مبتنی بر رفتار نرمال، بدون استفاده از داده‌های برچسب‌گذاری شده	E4	BPIC 2012, BPIC 2013, BPIC 2015, BPIC 2017	-	Junior & et al., (2020)
استفاده از اهرم آماری برای کشف دنباله‌های ناهنجار در محیط فرایندکاوی آنلاین. برای این کار نویسندگان از امتیاز ناهنجاری با توجه به اطلاعات رویداد، به عنوان معیاری جهت تشخیص ناهنجاری استفاده کردند.	E4	-	-	Ko & Comuzzi, (2020)
معرفی یک روش جدید برای کشف ناهنجاری‌ها به صورت بلادرنگ ^۳ در جریان رویدادها. ^۴ در این رویکرد برای کشف ناهنجاری‌ها، از شبکه‌های عصبی خودرمزگذار استفاده شده است که به عنوان یک فیلترکننده، اقدام به شناسایی و حذف ناهنجاری‌ها از جریان رویدادها در فرایندکاوی آنلاین می‌نماید. برای شناسایی ناهنجاری‌ها، از خطای میانگین مربع بین بردار ورودی و بردار خروجی شبکه عصبی استفاده می‌شود.	E4	BPIC 2019	-	Krajsic & (Franczyk, 2020)
ارائه روشی برای کشف نمونه فرایندهای ناهنجار، بدون نیاز به دانش قبلی در مورد فرایند. در این مقاله نویسندگان با استفاده از ترکیب شبکه‌های عصبی خودرمزگذار با گراف، مفهوم نوینی تحت عنوان «گراف خودرمزگذار» ایجاد نموده‌اند. برای این منظور آن‌ها ارتباطات بین فعالیت‌ها را به عنوان یال در نظر گرفته و آن را به یک گراف تبدیل کردند، تا بتوانند با استفاد از آن، رویدادهای ناهنجار را کشف کنند. از این گراف رمز شده برای فهم اطلاعات ساختاری فرایند استفاده می‌شود.	موردهای ناهنجار / پرت	BPIC 2012, BPIC 2013, BPIC 2017	-	Huo & et al., (2021)

<http://stlm.gom.ac.ir>

1. Simple Auto Encoder Based Anomaly Detection
2. Probabilistic Auto Encoder Based Anomaly Detection
3. Real time
4. Event streams

هدف	جنبه‌های کیفیت داده مورد تاکید	مجموعه داده استاندارد مورد استفاده	ابزار پیاده‌سازی	منبع
توسعه یک روش نظری اطلاعات ^۱ برای کشف ناهنجاری در گزارش‌های رویداد. در این مطالعه رویکردی جدیدی برخلاف دیگر رویکردها مانند تشخیص ناهنجاری‌های احتمالی مبتنی بر فاصله، یا بازسازی معرفی شده است. بدین ترتیب آن‌ها از مفهومی به نام اهرم آماری برای درک اینکه یک مقدار معین در چه فاصله‌ای از دیگر مقادیر قرار گرفته است، برای تشخیص ناهنجاری استفاده کرده‌اند. همچنین برای تعیین آستانه ناهنجاری، سه روش مختلف مورد تحلیل و بررسی قرار گرفته است.	C2, C3	BPIC 2012, BPIC 2013, BPIC 2017	Python	Ko & Comuzzi,) (2021
ارائه روشی بدون نظارت مبتنی بر مدل‌های مولد عمیق ^۲ مانند یک شبکه عصبی خودرمزگذار متغیر ^۳ (VAE)، برای کشف ناهنجاری‌ها و رفتارهای کم‌تکرار در داده‌های رویداد درون محیط‌های برخط، با هدف بهبود نتایج کشف فرایند	E4	BPIC 2013	-	Krajsic &) (Franczyk, 2021

جدول ۴- پژوهش‌های بررسی شده مربوط به دسته بازسازی و ترمیم

هدف	جنبه‌های کیفیت داده مورد تاکید	مجموعه داده استاندارد مورد استفاده	ابزار پیاده‌سازی	منبع
روشی برای بازسازی رویدادهای گمشده در گزارش رویداد، با استفاده از یک رویکرد احتمالی مبتنی بر شبکه‌های پتری تصادفی ^۴ و شبکه‌های بیزی ^۵	E5	-	-	Rogge-Solti &) (et al., 2013b
این مقاله مربوط به پژوهش قبلی نویسندگان (Rogge-Solti & et al., 2013b) است که بخشی از نتایج آن را در این گزارش ارائه داده‌اند. در این پژوهش با استفاده از دانش به دست آمده از مدل فرایند موجود در مستندات سازمان، مقادیر گمشده در یک گزارش رویداد، بازسازی شده است.	E5	-	-	Rogge-Solti &) (et al., 2013a
در این پژوهش، به صورت تئوری اثبات شده که چالش بازیابی، از نوع مسائل NP-Hard است. سپس یک روش بازیابی برای رویدادهای گمشده براساس پالایش شبکه پتری ^۶ ارائه می‌دهند.	E5	-	-	Wang & et al.,) (2013

1. Information-theoretic
2. Deep generative models
3. Variational autoencoder
4. Stochastic Petri nets
5. Bayesian networks
6. Petri net

منبع	ابزار پیاده‌سازی	مجموعه داده استاندارد مورد استفاده	جنبه‌های کیفیت داده مورد تاکید	هدف
Song & et al.,) (2015	-	-	Ti1, Ti4, Ti5	ارائه یک روش اکتشافی کارا براساس تجزیه فرایند، ^۱ به همراه یک الگوریتم اکتشافی برای حل چالش رویدادهای گم شده در گزارش رویداد. الگوریتم اکتشافی معرفی شده با کاهش فضای مسأله تلاش می‌نماید رویدادهای گمشده را پیدا کند.
Wang & et al.,) (2015	-	-	E4	در این پژوهش با استفاده از مفهوم گراف، ساختارهای معیوب ^۲ شناسایی و براساس ویژگی‌های ساختاری بین گروه‌ها، برچسب رویدادهای ناسازگار اصلاح می‌شود.
Xia & et al.,) (2016	ProM	-	E5	در این گزارش فنی، یک افزونه به نام Effa برای چارچوب نرم‌افزاری فرایندکاوی ProM ارائه گردید که براساس مشخصات فرایند، اقدام به بازیابی خودکار گزارش رویداد می‌نماید.
Song & et al.,) (2016	-	-	Ti3	بازسازی برچسب‌های زمانی با استفاده از محدودیت‌های زمانی ^۳ که حداقل و حداکثر انجام یک فعالیت را مشخص می‌نماید.
Wang & et al.,) (2016	-	-	E5	این مقاله در ادامه پژوهش قبلی نویسندگان (Wang et al., 2013) ارائه شده است. در این پژوهش یک چارچوب کلی برای بازیابی رویدادهای گمشده براساس محدودیت‌های کسب‌وکار و همچنین استفاده از مشخصات فرایند معرفی شده است.
Conforti &) (et al., 2018	ProM	-	Ti1	معرفی یک رویکرد خودکار برای اصلاح برچسب‌های زمانی دارای خطا، با هدف مرتب نمودن رویدادها
Dixit & et al.,) (2018	ProM	BPIC 2015, Sepsis	Ti1, Ti2	در این پژوهش سه شاخص برای حل مشکل مرتب کردن رویدادها براساس برچسب‌های زمانی معرفی شده است.
(Nguyen, 2018)	-	BPIC 2012, BPIC 2013	E5	ارائه روشی دو مرحله‌ای براساس شبکه عصبی خودرمزگذار برای کشف ناهنجاری و بازسازی مقادیر گمشده. بخشی از این پایان‌نامه در مقاله (Nguyen & Comuzzi, 2018) و مقاله (Nguyen & et al., 2019) منتشر شده است.
Sani & et al.,) (2018b	ProM RapidProM	BPIC 2012, BPIC 2013, BPIC 2017, Hospital Billing, Road Fines, Sepsis	E4	توسعه یک روش ترمیم (بازسازی) برای رویدادهایی که دارای رفتار پرت هستند، براساس روشی احتمالی برای تشخیص رویدادهای پرت با توجه به زمینه فرایند

<http://stlm.gom.ac.ir>

1. Process decomposition
2. Unsound
3. Temporal constraints

منبع	ابزار پیاده سازی	مجموعه داده استاندارد مورد استفاده	جنبه های کیفیت داده مورد تاکید	هدف
Kong & et al.,) (2019	java	-	T1	معرفی یک چارچوب سه مرحله ای برای بازسازی فعالیت های یک دنباله دارای خطا با استفاده از یافتن نزدیک ترین دنباله مشابه، با قابلیت پشتیبانی از ساختارهای حلقه، نقاط تصمیم گیری، فعالیت های همزمان و فعالیت های متوالی
Sim & et al.,) (2019	-	-	E5	توسعه یک رویکرد احتمالی با هدف جایگزینی مقادیر از دست رفته تکی و چندگانه، به وسیله تحلیل روابط بین رویدادها از طریق تعامل اشیاء کسب و کار و تراکنش های زمانی
Sani, van Zelst) (& et al., 2019	ProM RapidProM	BPIC 2012, BPIC 2013, BPIC 2017, BPIC 2018, Road Fines, Hospital Billing, Sepsis	A1, A5	ارائه یک رویکرد ترکیبی شامل پیش پردازش گزارش رویداد، برای کشف فعالیت های پرت و بازسازی آنها براساس زمینه ای که فعالیت ها در آن جریان دارد. برای این منظور از احتمالات شرطی برای جایگزینی مقادیر پرت با مناسب ترین جانشین استفاده می شود. همچنین برای تعیین رفتار پرت، از یک حد آستانه استفاده شده است.
Song & et al.,) (2019	ProM	-	T2	ارائه یک رویکرد چهار مرحله ای به نام Shel، برای ترمیم یک گزارش رویداد با دنباله های دارای انحراف
Xu & Liu,) (2019	-	-	A6	ارائه یک چارچوب جدید برای خوشه بندی دنباله ها و مطابقت دنباله های دارای فعالیت گمشده با خوشه های ایجاد شده در مرحله قبل، سپس بازسازی فعالیت های گمشده و در نهایت کاوش مدل فرایند براساس زیردنباله ها
Conforti &) (et al., 2020	Java	-	Ti1	ارائه روشی خودکار برای بازسازی ترتیب رویدادهای مربوط به یک نمونه فرایند، که زمان رخداد آن ها به صورت مشابه در لاگ ثبت شده باشد.
Denisov &) (et al., 2020	ProM	-	Ti3	در این مطالعه رویکردی جدید از مشکلات کیفیت داده در گزارش رویداد مطرح می شود. به علت اشتراک منابع و رقابت بین مشارکت کنندگان فرایند، ممکن است بعضی از مشارکت کنندگان به منبع دسترسی پیدا نکنند، یا رویدادها انجام شوند، اما در سامانه ثبت نگردند. برای حل این چالش از یک روش جدید جهت بازسازی گزارش های رویداد، با استفاده از برنامه ریزی خطی ^۱ ، جهت یافتن محدوده زمانی رویدادهای گمشده استفاده شده است.

منبع	ابزار پیاده‌سازی	مجموعه داده استاندارد مورد استفاده	جنبه‌های کیفیت داده مورد تاکید	هدف
Helal & Awad,) (2020	-	BPIC 2013	E2	بازسازی شناسه نمونه موردها برای رویدادهای بدون برچسب، از طریق دانش موجود مربوط به مدل فرایند، با پشتیبانی از فعالیت‌های دارای چرخه به صورت آنلاین (در زمان اجرا)
Horita & et al.,) (2020	Pthon, scikit-learn library	BPIC 2013	E5	در این پژوهش یک روش جدید برای جایگذاری مقادیر گمشده معرفی شده است. در این راستا، از یک درخت تصمیم استفاده شده است که با استفاده از آن می‌توان مقادیر گم شده را جایگزین نمود و همچنین قوانین لازم برای درک دلایل ایجاد مقادیر گم شده را استخراج نمود.
Sadeghianasl) (& et al., 2020	-	BPIC 2019	A1	ارائه یک رویکرد ترمیم و بازسازی گزارش رویداد براساس دیدگاه بازی‌وارسازی، که با هدف ترمیم برچسب فعالیت‌های دارای اشتباه توسعه یافته است. برای این منظور از اطلاعات جمع‌آوری شده از متخصصان این زمینه استفاده شده است.
Liu & et al.,) (2021	-	BBPIC 2012, Hospital Billing	A6	بازسازی فعالیت‌های گمشده براساس تحلیل رابطه توالی ^۱ بین فعالیت‌ها توسط یک ماتریس و سپس خوشه‌بندی همه دنباله‌ها. همچنین در ادامه، بازسازی دنباله‌های ناقص براساس شباهت به نزدیک‌ترین دنباله ممکن انجام شده است. برای خوشه‌بندی دنباله‌ها از روش شبکه‌های خودسازمانده (SOM) استفاده شده است.