مروری بر روش‌های مبتنی بر یادگیری ماشین در تولید لاگ رخداد مصنوعی

جلال رضایی نور * | دکترای مهندسی صنایع، استاد،
گروه مهندسی صنایع، دانشکده فنی و مهندسی، دانشگاه قم، قم،ایران

منصوره یاری ایلی | دکترای مهندسی فناوری اطلاعات، پژوهشگر پسادکتری
گروه مهندسی صنایع، دانشکده فنی و مهندسی، دانشگاه قم، قم،ایران

دریافت: ۱۴۰۵/۰۴/۳۰ | **پذیرش:** ۱۴۰۵/۰۶/۲۵ | مقاله برای اصلاح به مدت ۱۸ روز نزد پدیدآوران بوده است.

چکیده: کمبود داده‌های واقعی به دلیل محدودیت‌های حریم خصوصی، هزینه‌های جمع‌آوری و کیفیت پایین داده، توسعه و ارزیابی الگوریتم‌های فرایندکاوی را با چالش مواجه کرده است. تولید داده‌های مصنوعی که ضمن حفظ ویژگی‌های آماری و ساختاری فرایند، ارتباط مستقیمی با نمونه‌های واقعی نداشته باشند، راهکاری مؤثر برای رفع این محدودیت‌ها محسوب می‌شود. هدف این پژوهش، ارائه مروری نظام‌مند بر روش‌های تولید لاگ رخداد مصنوعی در حوزه فرایندکاوی است که بر پایه الگوریتم‌های یادگیری ماشین توسعه یافته‌اند. جستجوی نظام‌مند در پایگاه‌هایScienceDirect ، PubMed، IEEE Xplore، SpringerLink تا مرداد ۱۴۰۵ (آگوست ۲۰۲۵) صورت گرفت. از میان ۱۹۹۰ رکورد اولیه، پس از حذف مقالات تکراری و غربالگری در سه مرحله (عنوان، عنوان و چکیده، متن کامل) در نهایت ۱۵ مطالعه برای تحلیل نهایی انتخاب شدند.

یادگیری ماشین پارادایمی پیشرفته در تولید لاگ رخداد مصنوعی است و بسته به نوع روش توسعه داده شده و روند تکاملی از روش‌های یادگیری ماشین سنتی (مانند درخت تصمیم)، تا روش‌های پیشرفته‌تر مبتنی بر یادگیری عمیق و هوش مولد (مانند مدل‌های زبانی، اتوانکودرها و...) و روش‌های ترکیبی را در برمی‌گیرد. تحلیل مطالعات، نشان‌دهنده توجه فزاینده به دسته دوم یعنی روش‌های مبتنی بر یادگیری عمیق و هوش مولد است. هدف اصلی تولید لاگ‌های رخداد مصنوعی ارزیابی الگوریتم پنج مقاله (۳۳٪)، حفظ محرمانگی پنج مقاله (۳۳٪)، تولید داده منفی دو مقاله (۱۳٪) تولید داده چند بعدی دو مقاله (۱۳٪) تحلیل سناریو یک مقاله (۷٪) عنوان شده است. در تمام مقالات معیارهای وفاداری (یعنی سنجش شباهت آماری داده‌های مصنوعی به داده‌های واقعی) برای ارزیابی روش استفاده شده است، درحالی‌که کاربردپذیری در فرایندکاوی (یعنی ارزیابی روش‌های کشف یا بررسی انطباق با داده‌های مصنوعی) در هشت مقاله (۵۳٪) و محرمانگی تنها در دو مقاله (۱۳٪) گزارش شده است. فقدان چارچوب ارزیابی یکپارچه و نیز نبود ارزیابی‌های محرمانگی، چالش اصلی در مقایسه روش‌هاست. مسیرهای آتی پژوهش بر ترکیب توانایی‌های

نشریهٔ علمی (رتبه ین المللی)
پژوهشگاه علوم و فناوری اطلاعات ایران
شاپا(چاپی) ۸۲۲۳-۲۲۵۱
شاپا(الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISTA، SCOPUS و ISC
http://jipm.irandoc.ac.ir
دوره xx | شمارهٔ x | ص ص xx-xx
۱۳xx x

نوع مقاله: مروری

به این مقاله به شکل زیر استناد کنید:
درون متن:
(رضایی نور، ی‌ار‌ای‌لی، زودآیند)
در فهرست منابع:
رضایی نور، جلال. ی‌ار‌ای‌لی، منصوره. زودآیند. **مروری بر روش‌های مبتنی بر ی‌ادگی‌ری ماشی‌ن در تولی‌د لاگ رخداد مصنوعی**. *پژوهشنامه پردازش و مدیریت اطلاعات*.
http://Jipm.irandoc.ac.ir (دسترسی در روز/ماه/سال)

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

X xx ۱۳ | دورهٔ xx | شمارهٔ x

مدل‌های مولد با مدل‌های ساخت‌یافته (مانند شبکه‌های پتری) برای تضمین اعتبار ساختاری داده‌های تولیدشده از طریق ترکیب داده و دانش و نیز ارزیابی مقایسه‌ای مجموعه‌ای از مکانیسم‌های تولید (یعنی درک چگونگی تأثیر انتخاب مکانیسم‌های تولید مختلف بر عملکرد الگوریتم‌های فرایندکاوی) متمرکز خواهد بود. این سوال همچنان بی‌پاسخ مانده است که کدام مکانیسم(های) تولید، لاگ‌های رخداد بهتری برای فرایندکاوی تولید می‌کنند.

کلیدواژه‌ها: لاگ رخداد مصنوعی، فرایندکاوی، یادگیری ماشین، یادگیری عمیق، مقاله مروری

Email: j.rezaee@qom.ac.ir

۱. مقدمه

در سال‌های اخیر فرایندکاوی به‌عنوان یک حوزه نوظهور در تقاطع مدیریت فرایند، علم داده و هوش مصنوعی، نقش مهمی در استخراج دانش از داده‌های ثبت‌شده در سیستم‌های اطلاعاتی ایفا کرده است. هدف اصلی فرایندکاوی، کشف، پایش و بهبود فرایندهای تجاری از طریق تحلیل رخدادهای ثبت‌شده در سامانه‌هایی مانند برنامه‌ریزی منابع سازمانی[۱]، مدیریت زنجیره تأمین[۲] و مدیریت ارتباط با مشتری[۳] است (Yari Eili et al. 2023). هسته اصلی هر پروژه فرایندکاوی، لاگ رخداد است. لاگ رخداد مجموعه‌ای از فعالیت‌های انجام شده و مشخص برای یک نمونه فرایند (مانند یک بیمار) بوده و مجموعه فعالیت‌های مرتبط با یک نمونه فرایند، یک مسیر فرایند را تشکیل می‌دهند. برای استانداردسازی ذخیره‌سازی و تبادل این داده‌ها قالب XES[۴] توسعه یافته است(Acampora et al. 2017).

با وجود اهمیت کیفیت لاگ رخداد برروی عملکرد الگوریتم‌های فرایندکاوی، دسترسی به داده‌های واقعی همواره امکان‌پذیر نیست. در بسیاری از حوزه‌ها به‌ویژه حوزه‌های سلامت، مالی و دولتی، داده‌ها شامل اطلاعات حساس و محرمانه هستند که انتشار یا اشتراک‌گذاری آن‌ها متاثر از قوانین حفاظت از داده، نگرانی‌های امنیتی و محدودیت‌های مالکیت فکری خواهد بود (Kamal et al. 2024). حتی در شرایطی که داده‌های واقعی در دسترس باشند، جمع‌آوری و آماده‌سازی لاگ رخداد در قالب مناسب الگوریتم‌های فرایندکاوی، به دلیل وجود مسیرهای پیچیده،

[1] ERP
[2] SCM
[3] CRM
[4] eXtensible Event Stream

وابستگی‌های زمانی، منابع و ویژگی‌های متنوع، اغلب زمان‌بر و پرهزینه است (Salehi et al. 2023, Yari Eili et al. 2022). از طرف دیگر کیفیت پایین داده، وجود خطا، مقادیر گم‌شده و ثبت ناقص رخدادها می‌تواند به‌طور مستقیم بر عملکرد الگوریتم‌های کشف فرایند و بررسی انطباق، تأثیر بگذارد (Andrea et al. 2025).

از سوی دیگر توسعه و ارزیابی الگوریتم‌های جدید فرایندکاوی نیازمند مجموعه داده‌هایی با پارامترها و ویژگی‌های قابل کنترل است (Di Ciccio et al. 2015). اگرچه برخی لاگ‌های رخداد واقعی به‌صورت عمومی در دسترس هستند اما تعداد آن‌ها محدود بوده و معمولا فاقد تنوع کافی از نظر اندازه لاگ، طول مسیر، میزان خطا، پیچیدگی ساختاری و الگوهای رفتاری مختلف هستند. این محدودیت باعث می‌شود امکان ارزیابی مستقل جنبه‌های مختلف الگوریتم‌های فرایندکاوی، مانند مقیاس‌پذیری، مقاومت در برابر خطا و توانایی مدل‌سازی فرایندهای پیچیده کاهش یابد (Skydanienko et al. 2018). بنابراین نیاز به روش‌هایی که بتوانند داده‌های مصنوعی مبتنی بر واقعیت با پارامترها و ویژگی‌های قابل تنظیم تولید کنند احساس می‌شود.

در این راستا تولید لاگ رخداد مصنوعی به‌عنوان رویکردی موثر برای رفع محدودیت‌های داده‌های واقعی مطرح شده است. در این رویکرد با یادگیری الگوهای آماری و ساختاری موجود در داده‌های واقعی، نمونه‌های مصنوعی جدیدی تولید می‌شوند که ضمن حفظ ویژگی‌های کلیدی فرایند، ارتباط مستقیمی با افراد یا نمونه‌های واقعی ندارند. این ویژگی، امکان اشتراک‌گذاری امن داده‌ها، توسعه و ارزیابی الگوریتم‌های جدید و ایجاد محیط‌های آزمایشی کنترل‌شده را فراهم می‌کند. همچنین لاگ‌های مصنوعی می‌توانند با تعریف ویژگی‌هایی مانند تعداد نمونه، سطح نویز، طول مسیر متفاوت یا ساختارهای پیچیده فرایندی تولید شوند و این امر به پژوهشگران اجازه می‌دهد تا عملکرد روش‌های فرایندکاوی را در سناریوهای مختلف ارزیابی کنند (Jouck and Depaire 2019).

روش‌های اولیه تولید لاگ رخداد مصنوعی عمدتا مبتنی بر شبیه‌سازی مدل فرایند بودند، مانند مدل‌های تصادفی (Kuhn et al. 2024)، شبکه‌های پتری (Nedopetalski, de Freitas, 2021) و شبیه‌سازی رویدادگسسته (Friederich et al. 2026). با پیشرفت یادگیری ماشین و به‌ویژه ظهور مدل‌های مولد عمیق، رویکردهای جدیدی مانند شبکه‌های بازگشتی (RNN/LSTM)، شبکه‌های مولد تخاصمی(GAN) ، مدل‌های انتشار[1] ، مدل‌های زبانی بزرگ

[1] Diffusion Models

و روش‌های ترکیبی مبتنی بر شبیه‌سازی و یادگیری داده‌محور توسعه یافته‌اند (Fonseca and Bacao, 2023). این روش‌ها قادرند وابستگی‌های پیچیده زمانی، رفتاری و چندبعدی موجود در داده را یاد بگیرند و داده‌هایی با شباهت آماری بالا با داده‌های واقعی تولید کنند (Lautrup et al. 2024). علاوه بر آن، مسائل مربوط به حریم خصوصی منجر به توسعه روش‌هایی مانند یادگیری تفاضلی خصوصی[1] و مدل‌های مولد امن شده است که تلاش می‌کنند تعادل مناسبی میان سودمندی داده[2] و کاهش خطر افشای اطلاعات برقرار کنند (Goyal and Mahmoud, 2024).

با وجود رشد سریع روش‌های تولید لاگ رخداد مصنوعی و افزایش تنوع رویکردهای پیشنهادی، مطالعات موجود پراکنده بوده و طبق دانش نویسندگان، تاکنون بررسی تکنیک‌های مورد استفاده و تحلیل روندهای پژوهشی این حوزه کمتر مورد توجه بوده است. شناخت دقیق روش‌های موجود و نقاط قوت و محدودیت‌های آن می‌تواند مسیر توسعه نسل آینده ابزارهای تولید داده مصنوعی در فرایندکاوی را مشخص کند.

بنابراین مطالعه حاضر به مرور نظام‌مند روش‌های تولید لاگ رخداد مصنوعی مبتنی بر یادگیری ماشین می‌پردازد. این مطالعه با بررسی مطالعات منتشرشده، رویکردهای موجود را دسته‌بندی کرده، تکنیک‌ها و الگوریتم‌های مورد استفاده را تحلیل و روند تکامل این حوزه را مشخص می‌سازد. نتایج این پژوهش می‌تواند به پژوهشگران و توسعه‌دهندگان کمک کند تا درک جامع‌تری از جایگاه فعلی الگوریتم‌های یادگیری ماشین در تولید لاگ رخداد مصنوعی به دست آورده و مسیرهای تحقیق آتی را شناسایی کنند.

۲. پیشینه پژوهش

داده‌های ورودی با کیفیت بالا برای توسعه بهتر، درک نتایج و تصمیمات آگاهانه در رویکردهای مبتنی بر داده ضروری است. تضمین کیفیت سیستم‌های اطلاعاتی نیازمند مجموعه داده‌های بزرگی برای آزمایش سناریوهای واقع‌بینانه و اعتبارسنجی است (Malin and Goodman, 2018). متأسفانه در دسترس بودن داده‌های واقعی چالش‌های منحصر به فردی ایجاد می‌کند که مانع از اشتراک‌گذاری آزاد به دلیل نگرانی‌های مربوط به حریم خصوصی می‌شود. جدا از نگرانی‌های حریم خصوصی، داده‌های واقعی گاهی اوقات ممکن است به دلیل فقدان یک سیستم اطلاعاتی

[1] Differential Privacy
[2] Data Utility

برای ثبت سیستماتیک اطلاعات، اندازه کوچک مجموعه داده یا توزیع آماری مغرضانه که آزمایش یا آموزش الگوریتم‌های یادگیری ماشین را محدود می‌کند، در دسترس نباشند.

برای غلبه بر این مسائل، محققان در حال یافتن راه‌هایی برای انتشار امن مقدار داده‌ها به جامعه تحقیقاتی هستند. ابتکارات قبلی شامل تکنیک‌های پنهان‌سازی و ناشناس‌سازی داده‌ها برای تبدیل داده به روشی بود که ویژگی‌های آماری آن را حفظ می‌کرد و در عین حال مانع از هرگونه نشت حریم خصوصی می‌شد (Fahrenkrog-Petersen et.al, 2018). با این‌حال تکنیک‌های ناشناس‌سازی دارای نقص‌های زیاد و مستعد خطرات حریم خصوصی باقی‌مانده هستند، به ویژه هنگام برخورد با مقدار محدودی از داده‌های ثبت شده نامناسب هستند، همچنین داده‌های حاصل، بخش زیادی از صحت خود را از دست می‌دهند و کاربرد کمی برای تحقیقات پیشرفته دارند (Pawar et al. 2018).

تولید داده‌های مصنوعی به‌عنوان راه‌حلی مناسب ظهور کرده است و مقادیر قابل توجهی از داده‌های مصنوعی را برای آموزش مدل فراهم می‌کند. روبین در سال ۱۹۹۳ مفهوم داده‌های مصنوعی را معرفی کرد و استفاده از جایگذاری‌های چندگانه بر روی همه متغیرها را پیشنهاد داد به طوری‌که هیچ یک از داده‌های اصلی منتشر نشود (Rubin 1993). داده مصنوعی به شبیه‌سازی توزیع مجموعه داده‌ها از طریق یک فرآیند آماری برای استخراج اطلاعات از یک مجموعه داده واقعی اشاره دارد که برای مصرف عمومی تولید می‌شود. مجموعه داده‌های مصنوعی دارای همان ویژگی‌های آماری اصلی است و در عین حال مقاومت بهتری در برابر حملات حریم خصوصی نشان می‌دهد (Bellovin et al. 2019). این مسئله با روش‌های افزایش داده، که در آن داده‌های موجود را دستکاری می‌کنند، متمایز است، زیرا داده مصنوعی با نمونه‌برداری از توزیع‌های آموخته شده، داده‌های متنوع جدیدی تولید می‌کند. داده‌های مصنوعی زمانی ارزشمند می‌شوند که داده‌های واقعی ناکافی، برچسب‌گذاری داده‌ها پرهزینه یا داده‌ها دارای توزیع‌های مغرضانه باشند. داده‌های مصنوعی مزایای متعددی به همراه دارند (James et al. 2019): مقاومت در برابر حملات حریم خصوصی، جایگزینی ارزان برای داده‌های واقعی، امکان تولید نامحدود نمونه‌های واقعی خاص مطابق با الزامات، محدودیت‌های نظارتی کمتر، انتشار سریع‌تر و در نتیجه زمان کوتاه‌تر برای دستیابی به بینش‌های مبتنی بر داده. هنگام تولید داده‌های مصنوعی دو هدف وجود دارد که به رقابت می‌پردازند: سودمندی بالای داده‌ها (یعنی اطمینان از مفید بودن داده‌های

مصنوعی با توزیعی نزدیک به داده‌های اصلی) و ریسک افشای پایین.
ایجاد تعادل بین این دو هدف می‌تواند دشوار باشد زیرا افزایش محرمانگی با هزینه‌هایی در سودمندی همراه است.

در فرایندکاوی، اولین تلاش در سال ۲۰۰۵ برای تولید لاگ رخداد مصنوعی برای ارزیابی تکنیک‌های فرایندکاوی با استفاده از ابزارهای CPN و شبیه‌سازی شبکه‌های پتری رنگی انجام شد (De Medeiros and Günther, 2005). با تکیه بر این پیشرفت، روش‌های مبتنی بر شبیه‌سازی مانند روش‌های رویه‌ای در جهات مختلف تکامل یافتند. به موازات آن، تولید لاگ رخداد مصنوعی با استفاده از پیمایش مبتنی بر اتوماتا یا حل‌کننده‌هایی مانند Alloy، SAT و ASP به سمت روش‌های اعلانی و مبتنی بر محدودیت گسترش یافته است. اخیرا تمرکز بیشتری برروی استفاده از روش‌های داده‌محور بوده است. این روش‌ها از LSTMها، GANها، CVAEها و LLMها برای یادگیری الگوهای پیچیده فرایند از لاگ رخداد واقعی استفاده می‌کنند.

۳. روش تحقیق

پژوهش حاضر به‌منظور شناسایی، دسته‌بندی و تحلیل نظام‌مند مطالعات مرتبط با تولید لاگ رخداد مصنوعی در فرایندکاوی که مبتنی بر الگوریتم‌های یادگیری ماشین توسعه یافته‌اند و براساس چارچوب مرور نظام‌مند ارائه‌شده توسط (Xiao and Watson, 2019) انجام شده است. این چارچوب شامل تدوین پروتکل پژوهش، تعریف پرسش‌های تحقیق، جستجوی نظام‌مند منابع، غربالگری مطالعات، استخراج داده‌ها و تحلیل یافته‌ها است. در این مطالعه، ابتدا پرسش‌های پژوهش تدوین شد و سپس فرآیند جستجو، انتخاب و تحلیل مقالات مطابق با پروتکل مرور نظام‌مند اجرا گردید.

۱.۳ پرسش‌های تحقیق

هدف اصلی این مطالعه، ارائه تصویری جامع از روش‌ها، ابزارها، کاربردها و شیوه‌های ارزیابی تولید لاگ رخداد مصنوعی در حوزه فرایندکاوی است. بر این اساس، پرسش‌های پژوهش به صورت زیر تعریف شدند:

پرسش اول: چه روش‌ها، الگوریتم‌ها و رویکردهای مبتنی بر یادگیری ماشین برای تولید لاگ رخداد مصنوعی در فرایندکاوی ارائه شده‌اند؟

پرسش دوم: مهم‌ترین کاربردهای لاگ رخداد مصنوعی در مطالعات فرایندکاوی چیست؟

پرسش سوم: مطالعات مختلف کیفیت، سودمندی و کارایی لاگ رخداد مصنوعی را با استفاده از چه معیارهایی ارزیابی کرده‌اند؟

۲.۳ راهبرد جستجو

به‌منظور شناسایی مطالعات مرتبط، جستجوی نظام‌مند در پایگاه‌های علمی ScienceDirect، PubMed، IEEE Xplore و SpringerLink انجام شد. این پایگاه‌ها به دلیل پوشش گسترده پژوهش‌های حوزه فرایندکاوی و همچنین استفاده در مطالعات مروری پیشین انتخاب شدند. علاوه بر این مقالات شاخص منتشرشده در arXiv نیز مورد بررسی قرار گرفتند. جستجو کلیه مقالات منتشرشده تا مرداد ۱۴۰۵ (آگوست ۲۰۲۵) را دربر گرفت و محدودیتی از نظر سال انتشار اولیه اعمال نشد. جستجو در هر پایگاه به‌صورت مستقل و با استفاده از قابلیت جستجوی پیشرفته همان پایگاه انجام شد. عبارت جستجوی مورد استفاده به صورت زیر تعریف شد:

("process mining") AND ("synthetic event log generation" OR "artificial event log generation") AND ("machine learning")

این عبارات در چکیده و عنوان مقالات مرتبط به‌طور معمول مورد استفاده قرار می‌گیرند. استفاده از عملگر OR میان عبارت‌های *Synthetic Event Log Generation* و *Artificial Event Log Generation* به این دلیل بود که برخی مطالعات از اصطلاح دوم برای اشاره به داده‌های مصنوعی استفاده کرده‌اند. جستجوی اولیه در مجموع ۱۹۹۰ رکورد را بازیابی کرد که شامل ۱۲ مقاله از ScienceDirect، ۶۳۳ مقاله از PubMed ، ۹۳ مقاله از IEEE Xplore و ۱۲۵۲ مقاله از SpringerLink بود. پس از حذف مقالات تکراری ۷۹۶ مطالعه یکتا برای مرحله غربالگری باقی ماند. مدیریت و غربالگری منابع با نرم‌افزار EndNote انجام شد.

۳.۳ معیارهای ورود و خروج

پس از استخراج مطالعات مقالات براساس معیارهای ورود و خروج از پیش تعیین‌شده ارزیابی شدند.

معیار ورود

- مقالات نوشته شده به زبان فارسی و انگلیسی و داوری‌شده که به ارائه روش‌ها یا الگوریتم‌های تولید لاگ رخداد مصنوعی در فرایندکاوی پرداخته باشند. توسعه الگوریتم صرفا مبتنی بر روش‌های یادگیری ماشین باشد.

معیارهای خروج

- تمرکز اصلی مقاله بر تولید داده مصنوعی در فرایندکاوی نباشد و/یا روش تولید لاگ رخداد مصنوعی مبتنی بر یادگیری ماشین نباشد (برای مثال مبتنی بر شبیه‌سازی یا برنامه نویسی منطقی و... باشد).
- مقاله به زبانی غیر از انگلیسی و فارسی منتشر شده باشد.
- مقاله نسخه تکراری یک مطالعه دیگر باشد (مانند نسخه کنفرانسی که بعدا به مقاله مجله تبدیل شده است).

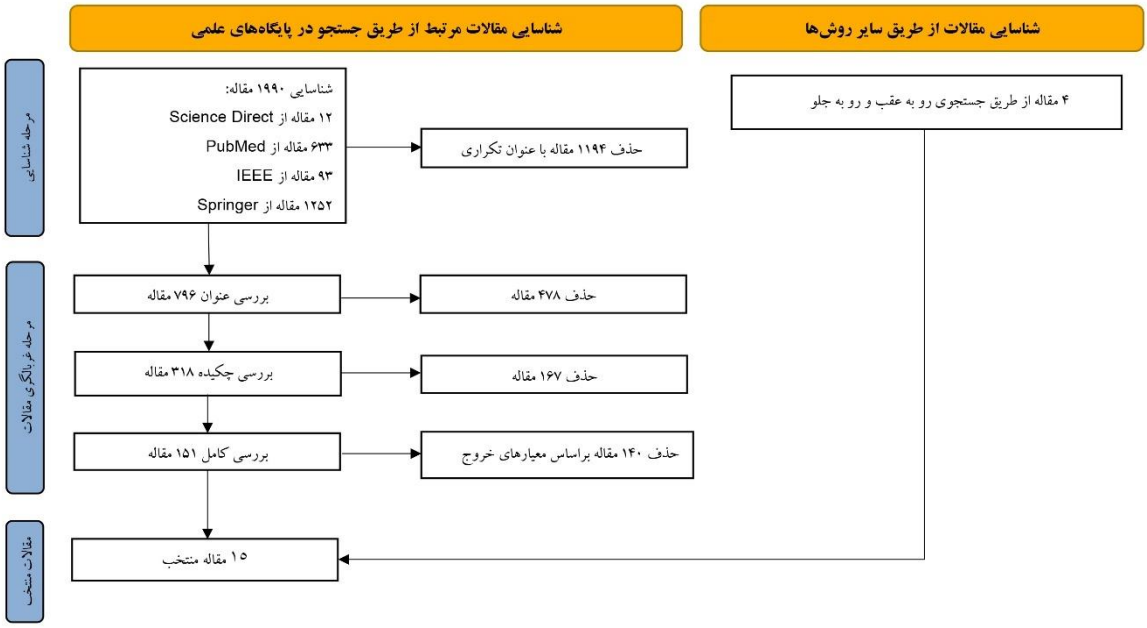
٤.٣ فرآیند غربالگری مطالعات

غربالگری مطالعات در سه مرحله انجام شد: در مرحله نخست مقالات تنها بر اساس عنوان بررسی شدند. در مرحله دوم عنوان و چکیده مقالات ارزیابی شد و در مرحله سوم متن کامل مقالات واجد شرایط مورد بررسی قرار گرفت. در مواردی که اطلاعات موجود برای تصمیم‌گیری کافی نبود، مقاله به مرحله بعدی غربالگری منتقل می‌شد. در مرحله نخست ۴۷۸ مقاله حذف و ۳۱۸ مقاله وارد مرحله بررسی عنوان و چکیده شدند. در این مرحله ۱۶۷ مقاله دیگر کنار گذاشته شد و ۱۵۱ مقاله برای ارزیابی متن کامل انتخاب گردید. به‌منظور افزایش جامعیت مرور، از دو روش جستجوی پسرو[1] از طریق بررسی منابع مقالات منتخب و جستجوی پیشرو[2] از طریق بررسی مقالات استنادکننده استفاده شد. با اعمال مجدد معیارهای ورود و خروج ۴ مطالعه دیگر شناسایی و به مجموعه نهایی افزوده شد. در نهایت ۱۵ مطالعه وارد مرور نظام‌مند شدند.

۵.۳ نمودار PRISMA

به‌منظور نمایش شفاف فرآیند انتخاب مطالعات، نمودار PRISMA در شکل ۱ ارائه شده است. این نمودار مراحل مختلف شناسایی، حذف موارد تکراری، غربالگری، ارزیابی متن کامل و انتخاب نهایی مطالعات را نشان می‌دهد و روند چندمرحله‌ای مرور نظام‌مند را مطابق با استانداردهای PRISMA مستند می‌سازد.

[1] Backward Search
[2] Forward Search



شکل ۱- نمودار فلوچارت این مقاله مروری

۴. نتایج

در این بخش، مقالات مرتبط براساس پرسش‌های تحقیقاتی بررسی و تحلیل می‌شوند.

۱.٤ پرسش اول: چه روش‌ها، الگوریتم‌ها و رویکردهای مبتنی بر یادگیری ماشین و یادگیری عمیق برای تولید لاگ رخداد مصنوعی در فرایندکاوی ارائه شده‌اند؟ و روند تحقیقاتی چگونه است؟

برای پاسخ به این پرسش ابتدا مقالات را از نظر روش توسعه داده شده دسته‌بندی می‌کنیم. برخلاف ابزارهای مبتنی بر شبیه‌سازی سنتی که به قوانین صریح و توزیع‌های احتمال تعریف‌شده توسط کاربر متکی هستند، رویکردهای مبتنی بر یادگیری ماشین و یادگیری عمیق از معماری‌های داده‌محور برای یادگیری خودکار ساختار بازسازی‌شده، توزیع‌های پیچیده زمانی، رفتاری و چندبعدی داده‌ها از لاگ رخداد استفاده می‌کنند. این رویکردها الگویی پیشرفته در تولید لاگ رخداد مصنوعی را نشان می‌دهند که طیف گسترده‌ای از مدل‌ها را شامل می‌شود؛ از روش‌های سنتی یادگیری ماشین (مانند درخت‌های تصمیم، مدل‌های مخفی مارکوف، شبکه‌های بیزی) گرفته تا روش‌های پیشرفته‌تر یادگیری عمیق و هوش مصنوعی مولد (مانند مدل‌های زبانی بزرگ، خودرمزگذار متغیر (VAE)، شبکه مولد تخاصمی (GAN) و مدل‌های احتمالی انتشار نویز (DDPMها)).

رویکردهای مبتنی بر یادگیری ماشین

با کار مستقیم بر روی گزارش‌های رویداد تاریخی برای استخراج قوانین زمانی، (Broucke et al., 2012) کشف فرآیند که یک روش غیرنظارتی است را به طبقه‌بندی نظارت‌شده تبدیل می‌کند. این رویکرد مبتنی بر چارچوب تولید رخداد منفی مصنوعی در AGNEsMiner است. نویسندگان به بررسی تعادل بین درستی (یعنی جلوگیری از رخدادهای منفی کاذب) و کامل بودن (یعنی القای رخدادهای منفی غیربدیهی) تحت ساختارهای پیچیده فرآیند می‌پردازند. نویسندگان معیارهای اطمینان را برای محدودیت‌های زمانی بازتعریف می‌کنند تا سوگیری فراوانی وقوع را از بین ببرند. یک تکنیک محاسبه واریانت حلقه، از عدم تطابق طول توالی رخداد ناشی از رفتار بازگشتی جلوگیری می‌کند. برای مدیریت طول‌های متغیر مسیر و پیچیدگی نمایی حلقه‌ها، مقایسه‌گر پنجره‌ای پویا، مقایسه‌های دقیق مبتنی بر موقعیت را با هم‌ترازی‌های توالی رخداد جایگزین می‌کند. در نهایت، مقاله یک رویکرد تولید مبتنی بر وابستگی غیر پنجره‌ای را تعریف می‌کند که رخدادهای منفی را صرفا از وابستگی‌های ساختاری، محلی و دور از دسترس استخراج می‌کند.

مدل‌های مخفی مارکوف (HMM) نوعی بسط داده شده از زنجیره‌های مارکوفی هستند که در آن‌ها توالی حالت زنجیره مارکوف پنهان است و هر حالت پنهان دارای احتمال انتشار برای نمادهای قابل مشاهده است. این مدل می‌تواند بسیاری از مسائل را با پیچیدگی کمتر در مقایسه با زنجیره‌های مارکوف ساده به تصویر بکشد. با این حال HMMها فاقد معناشناسی زمانی هستند. (Yang et. al, 2017) با گنجاندن برچسب‌های زمانی شروع و پایان فعالیت به طور مستقیم در گذارهای حالت، HMMهای سنتی را توسعه داده‌اند. این روش زمانی که منطق شفاف و ارتباط بالینی در اولویت هستند، به خوبی کار می‌کنند، اگرچه هنگام مدیریت داده‌های پیچیده یا نامنظم، نسبت به مدل‌های یادگیری عمیق، مقیاس‌پذیری و انعطاف‌پذیری کمتری دارند.

شبکه بیزی یک گراف غیرمدور جهت‌دار است که در آن متغیرهای تصادفی، گره‌ها و وابستگی‌های بین این متغیرها یال‌ها هستند. یال‌های جهت‌دار از گره والد به سمت گره فرزند امتداد دارند و وابستگی شرطی گره‌ها را تعریف می‌کنند. هر متغیر تصادفی دارای یک تابع توزیع احتمال پیوسته یا گسسته است. GEDI از بهینه‌سازی بیزی برای تولید SEهایی با ویژگی‌های متای هدف چندبعدی خاص و ارضاکننده استفاده می‌کند (Maldonado et al. 2024). این روش به

صورت تکراری، ابرپارامترهای یک مولد درخت فرآیند را با گسسته‌سازی یک فضای ویژگی چندبعدی، شامل نسبت‌های آماری و معیارهای آنتروپی گراف استخراج شده از طریق چارچوب FEEED، در یک شبکه مختصات، تنظیم می‌کند. GEDI فاصله اقلیدسی بین مقادیر ویژگی هدف و تولید شده را به حداقل می‌رساند و امکان ایجاد معیارهای عمدی را بدون نیاز به لاگ‌ها یا مدل‌های دنیای واقعی فراهم می‌کند. پارامترهای قابل تنظیم شامل ویژگی‌های هدف، پارامترهای مولد (مانند تعداد فعالیت‌ها، تعداد مسیرها و احتمال الگوهای ساختاری) و محدودیت‌های بهینه‌سازی هستند. انتخاب خودسرانه ویژگی‌ها می‌تواند منجر به عدم همگرایی بهینه‌سازی بیزی شود. همچنین با افزایش تعداد ویژگی‌ها پیچیدگی محاسباتی بالا می‌رود. این روش مبتنی بر پایتون توسعه یافته و کد منبع آن در گیت هاب در دسترس است.

روش‌های مبتنی بر هوش مصنوعی مولد

مدل‌های خودرگرسیو[1] مانند مبدل‌ها[2]، شبکه‌های عصبی بازگشتی و کانولوشنال داده‌ها را به‌صورت گام‌به‌گام و با تجزیه توزیع مشترک به مجموعه‌ای از احتمال‌های شرطی تولید می‌کنند. این رویکردها وابستگی‌های زمانی و ترتیب علّی رخدادها را حفظ می‌کند و ازاین‌رو برای توالی‌های رخداد و مسیرهای طولی مناسب هستند. در هر گام زمانی، این مدل‌ها با استفاده از حلقه‌های بازخورد، توزیع احتمال رخداد بعدی را با شرط قرار دادن یک تکه از مسیر[3] پیش‌بینی می‌کنند. برای تولید مسیرهای کامل، مدل به‌صورت خودرگرسیو از توزیع احتمال پیش‌بینی‌شده نمونه‌برداری می‌کند. این فرایند از یک توکن شروع آغاز شده و تا رسیدن به توکن پایان ادامه می‌یابد. این مدل‌ها برای مدیریت ورودی‌های با ابعاد بالا از تکنیک‌هایی مانند categorical embeddings، n-gramها و مقیاس‌بندی زمانی نسبی[4] استفاده می‌کنند و می‌توانند با استفاده از نزول گرادیان تصادفی با حفظ حریم خصوصی تفاضلی (DP-SGD) آموزش داده شوند تا تضمین‌های رسمی حریم خصوصی حفظ شود. روش‌های مبتنی بر Transformer انعطاف‌پذیری بالایی دارند و در حفظ شباهت توالی‌ها عملکرد قدرتمندی نشان می‌دهند، اما معمولا به تنظیمات دقیق و منابع محاسباتی قابل‌توجهی نیاز دارند و ممکن است در مواجهه با توالی‌های طولانی یا در حفظ حریم خصوصی با چالش مواجه شوند(Bauer et al. 2024).

[1] Autoregressive
[2] Transformers
[3] Prefix
[4] relative time scaling

مدل‌های انتشار[1] نسل جدیدی از روش‌های مولد هستند که از طریق مجموعه‌ای از مراحل، نویز را به‌تدریج به داده‌های ساختاریافته تبدیل می‌کنند. این مدل‌ها در کاربردهای فرایندکاوی برای تولید ساختارهای چندبعدی و مسیرهای موازی از نویز تصادفی خالص، عملکرد خوبی نشان داده‌اند. روش‌های انتشار، شباهت بالایی با داده‌های واقعی ایجاد می‌کنند و در کاربردهای حساس به حریم خصوصی یا کاربردهای با نیاز به دقت بالا در حال پیشی گرفتن ازGAN ها هستند. بااین‌حال از نظر محاسباتی پرهزینه بوده و تفسیرپذیری کمتری دارند (Lautrup et al. 2024).

(Camargo et al. 2021) در مقایسه تجربی میان رویکردهای مبتنی بر شبیه‌سازی و رویکردهای مبتنی بر یادگیری عمیق انجام داده‌اند. رویکرد شبیه‌سازی از Simod استفاده می‌کند که شامل کشف فرایند با SplitMiner، انطباق مسیر، استخراج پارامترهای مربوط به منابع، شاخه‌ها و زمان، و شبیه‌سازی با Bimp است. رویکرد یادگیری عمیق مبتنی بر معماری LSTM/GRU است که در آن لاگ رخداد به توالی‌های n-gram از فعالیت‌ها و زمان‌های پیوسته تبدیل می‌شوند. ویژگی‌های طبقه‌ای با لایه‌های embedding کدگذاری شده و برچسب‌های زمانی نرمال‌سازی یا لگاریتمی می‌شوند. مولد، توکن‌ها و زمان‌های نسبی بعدی را به‌صورت خودرگرسیو پیش‌بینی می‌کند و به‌جای استفاده از argmaxقطعی، از نمونه‌برداری تصادفی از توزیع‌های احتمال خروجی استفاده می‌کند تا از ایجاد حلقه جلوگیری کرده و تنوع رفتاری را تضمین کند.

برای رفع مشکلات ناپایداری و فروپاشی مُد که معمولا درGANها مشاهده می‌شود، PALSYN از لایه‌های LSTM/GRU یک‌جهته یا دوجهته برای مدل‌سازی وابستگی‌های ترتیبی و پیش‌بینی توزیع احتمال توکن بعدی استفاده می‌کند (Kuhn et al. 2025). مسیرها توکن‌سازی شده و از طریق لایه‌های embedding و masking پردازش می‌شوند و شبکه بازگشتی پس‌انتشار خطا و DP-SGD برای برش گرادیان و افزودن نویز آموزش می‌بینند. در نمونه‌برداری خودرگرسیو، مسیرهای جدید با پیش‌بینی احتمالی توکن‌های بعدی تولید می‌شوند. برای مواجهه با ویژگی‌های عددی و زمانی لاگ رخداد، ابزار PALSYN از الگوریتم DP-K-Means، یک خوشه‌بند فضایی چندبعدی و مبتنی بر سازوکار هندسی، برای گسسته‌سازی برچسب‌های زمانی پیوسته استفاده می‌کند. برای حفظ توازن میان حریم خصوصی و مطلوبیت، این روش‌های هندسی به

[1] Diffusion

تنظیم دقیق ابرپارامترهایی مانند تعداد خوشه‌ها وابسته هستند؛ زیرا تغییر فراوانی خوشه‌ها می‌تواند توزیع‌های آماری لاگ رخداد تولیدشده را دچار اعوجاج کند.

مدل‌های مبتنی بر خودرمزگذار[1]، داده‌های زمانی مصنوعی را با فشرده‌سازی رکوردها در نمایش‌های نهفته و سپس بازسازی آن‌ها تولید می‌کنند. گونه‌هایی مانند VAE و CVAE امکان کنترل انعطاف‌پذیر الگوهای زمانی و ویژگی‌های نمونه فرایند را فراهم می‌کنند.

خودرمزگذارهای واریشنال شرطی (CVAE) به تولید مسیرهای فرایندی چندبعدی[2] با شرطی‌سازی رمزنگار و رمزگشا برروی ویژگی‌های فرایند می‌پردازند. در مطالعه انجام شده توسط (Graziosi et al. 2025) یکCVAE ، مسیرهای فرایند که شامل داده‌های کنترل جریان ، برچسب زمانی و منابع هستند راتولید می‌کند. رمزنگار، یک LSTM را، برای رخدادهای ترتیبی، با لایه‌های کاملا متصل، برای ویژگی‌های ایستا، ترکیب کرده و ورودی‌ها را همراه با متغیر شرطی به یک فضای نهفته گاوسی نگاشت می‌کند. رمزگشا از سه LSTM خودرگرسیو برای پیش‌بینی فعالیت، برچسب زمانی و منبع بعدی استفاده می‌کند. مدل با استفاده از تابع زیان بازسازی cross-entropyو MSE و با واگرایی KL به‌صورت انتهابه‌انتها آموزش داده می‌شود. مسیرهای جدید با نمونه‌برداری گاوسی تحت شروط مشخص تولید می‌شوند و امکان شبیه‌سازی کنترل‌شده سناریوهایی مانند رانش مفهومی را فراهم می‌کنند. اگرچه خودرمزگذارها انعطاف‌پذیر و تفسیرپذیر هستند، اما آموزش آن‌ها ممکن است دشوار باشد و در صورت عدم سازگارسازی، برای داده‌های غیرساخت یافته مناسب نباشند.

GANها وDDPM ها چارچوب‌های مولد پیشرفته و غیرخودرگرسیوی هستند که برای یادگیری توزیع‌های پیچیده و با ابعاد بالا در لاگ رخداد طراحی شده‌اند. GAN از یک بازی کمینه-بیشینه خصمانه استفاده می‌کند که در آن مولد، مسیرهای مصنوعی را از نویز تصادفی ایجاد می‌کند تا تمایزگر[3] را فریب دهد. در مقابل، تمایزگر برای تشخیص مسیرهای واقعی از مسیرهای تولیدشده، آموزش می‌بیند. از سوی دیگر DDPMها تولید مسیر را به‌صورت یک فرایند حذف نویز تکرارشونده روی یک زنجیره مارکوف مدل می‌کنند و با یادگیری معکوس‌سازی فرایند انتشار پیشرو که به‌تدریج بردارهای لاگ رخداد را با نویز گاوسی مختل می‌کند، به تولید داده می‌پردازند. این مدل‌ها به‌ویژه در ثبت وابستگی‌های ساختاری بلندبرد و مسیرهای اجرای موازی

[1] Autoencoder-based
[2] Multi perpspective
[3] Discriminator

موثر هستند و می‌توانند با DP-SGDترکیب شوند تا از توزیع واریانت‌ها نمونه‌برداری کنند بدون آنکه داده‌های حساس کاربران افشا شود.

ProcessGANیک چارچوب مدل‌سازی خصمانه آگاه از زمان[1] مبتنی برGAN های شرطی است که از یک مولد مبتنی بر مبدل، یک متمایزگر مبتنی بر خود-توجه آگاه از زمان، و یک مولد مفهومی برای تولید فراداده‌هایی مانند اطلاعات دموگرافیک بیماران تشکیل شده است (Li et al. 2024). مولد از خودتوجه چندسر و رمزگشاهای مجزا استفاده می‌کند تا احتمال فعالیت‌ها و اختلاف‌های زمانی را به‌صورت موازی و با استفاده از نویز تصادفی و مدت‌زمان نمونه فرایند به‌عنوان ورودی‌های شرطی تولید کند. تمایزگر، اصالت توالی را با مدل‌سازی وابستگی‌ها میان فعالیت‌ها، برچسب‌های زمانی و فاصله‌های زمانی ارزیابی می‌کند. درحالی‌که فرایند آموزش، تابع زیان خصمانه را با زیان‌های کمکی MSE برای توزیع‌های فعالیت و زمان ترکیب می‌کند.

TraVaGو DDPM چارچوب‌های مولد عمیق و محافظ حریم خصوصی برای تولید لاگ رخداد مصنوعی هستند که تضمین‌های رسمی حریم خصوصی تفاضلی[2] ارائه می‌کنند (Wangelik et al. 2025). TraVaGمسیرهای فرایند را (که به‌صورتone-hot رمزگذاری شده‌اند) با یک GAN برای تولید مسیرهایی از نظر آماری سازگار، ترکیب می‌کند. در این چارچوب رمزگشا و تمایزگر با DP-SGD بهینه می‌شوند. DDPM از رویکرد مبتنی بر انتشار استفاده می‌کند که در آن نویز گاوسی طی یک فرایند مارکوف پیشرو اضافه شده و سپس طی یک فرایند معکوس مبتنی بر شبکه عصبی حذف می‌شود تا مسیرهای مصنوعی بازسازی شوند. این مدل نیز با DP-SGD آموزش داده می‌شود. در مقابل، رویکردهای مبتنی بر اغتشاش[3]، نویز لاپلاس را مستقیما و بدون آموزش مدل به مسیرها اضافه می‌کنند. این روش‌ها معمولا زمان اجرای کمتر از یک ثانیه داشته و به ابرپارامترهای کمی نیاز دارند. بااین‌حال، روش‌های سنتی مبتنی بر لاپلاس، مسیرهای کم‌تکرار را حذف می‌کنند که این امر موجب کاهش تنوع رفتاری و ایجاد احتمالی مسیرهای مصنوعی نامعتبر می‌شود. روش‌های مولد جدیدتر مبتنی بر DP-SGD، به‌جای اغتشاش مستقیم خروجی‌ها، نویز را در طول آموزش مدل وارد می‌کنند و در نتیجه، بدون حذف مبتنی بر فراوانی، امکان حفظ بهتر تنوع رفتاری را تحت تضمین‌های سخت‌گیرانه حریم خصوصی فراهم می‌سازند. بااین‌حال TaraVaG همچنان از هرس مبتنی بر آستانه استفاده می‌کند

[1] Time-Aware Adversarial Modeling
[2] Differential Privacy
[3] perturbation-based

که مسیرهای نادر را حذف کرده و در نتیجه تنوع رفتاری لاگ رخداد مصنوعی را به‌طور قابل‌توجهی کاهش می‌دهد.

مطالعه انجام شده توسط (Singh et al. 2024) مقایسه‌ای از مدل‌های مولد عمیق، یعنی مدل‌های بازگشتی در مقابل مدل‌های خصمانه، برای تولید داده از طریق بازمهندسی ویژگی‌ها ارائه می‌کند. در این مطالعه دو مدل مولد ارزیابی شده‌اند:

۱. یک مدل مبتنی بر LSTMکه از embedding رخدادها برای بازتولید با وفاداری بالا استفاده می‌کند؛
۲. یک GAN مبتنی بر مبدل که با استفاده از خودتوجهی، Gumbel-Softmaxو تابع زیان MSE، وابستگی‌های بلندبرد را ثبت کرده و مسیرهای متنوع تولید می‌کند.

دسته مدل‌های زبانی بزرگ و هوش مصنوعی مولد با بهره‌گیری از درک معنایی و مدل‌سازی توالی، تولید خودکار مصنوعات فرایندی، از جمله مدل‌های فرایند، لاگ رخداد و مسیرها را توسعه داده است. این رویکردها محدودیت‌های روش‌های شبیه‌سازی سنتی، از جمله کمبود مجموعه‌داده‌های چندبعدی در دسترس عموم، نیاز به قواعد دست‌ساز و زمان‌بر برای مدل‌سازی و دشواری مدیریت رفتارهای پیچیده فرایند را برطرف می‌کنند. مدل‌های زبانی بزرگ انعطاف‌پذیر هستند و می‌توان آن‌ها را از طریق تولید مبتنی بر دستور[1] به‌سرعت با زمینه‌های مختلف فرایندهای تجاری سازگار کرد، بدون آنکه به بازآموزی گسترده نیاز باشد. از نظر روش‌شناختی، این چارچوب‌ها معمولا بر مهندسی دستور یا معماری‌های ترکیبی متکی هستند که در آن یک مدل مولد احتمال رخداد بعدی را پیش‌بینی می‌کند و یک لایه اجرای رسمی، فضای جستجو را به رفتارهایی محدود می‌سازد که با قواعد فرایند مطابقت دارند (Long et al. 2024).

Terpsichoraاز GPT-4o برای خودکارسازی تولید مدل‌های فرایندی چندبعدی MP-Declare استفاده می‌کند و روابط پیچیده چندبعدی میان عناصر فرایند را که پیش‌تر در مخازن عمومی کمیاب بودند، ثبت می‌کند (da Silva Santos et al. 2024). یکی از نوآوری‌های کلیدی این روش، تولید توضیحات قواعد به زبان طبیعی است که تشخیص ناهنجاری معنایی را تسهیل می‌کند. ProcessTBench از GPT-4 برای تولید توالی‌های فراخوانی ابزار از پرس‌وجوهای متنی چندزبانه، از یک LLM Plan Variants Generatorبرای ایجاد

[1] prompt

راه‌حل‌های جایگزین، و از یک Event Log Parser برای تبدیل برنامه‌های متنی به لاگ رخداد استفاده می‌کند. برنامه‌های مبنا[1] که در ابتدا به‌صورت DAG نمایش داده شده‌اند، برای بررسی انطباق به Petri Netتبدیل می‌شوند (Redis et al. 2024).

FSM-GFlowNetsماشین‌های حالت متناهی را با شبکه‌های جریان مولد یکپارچه می‌کند تا لاگ‌های رخداد متنوعی از تعاملات کاربر با رابط کاربری تولید کند (Samanta 2025). پس از اینکه مسیرهای مبنا با استفاده از GPT-4o ساخته شد برای استخراج دانش، قواعد و محدودیت‌ها از طریق مهندسی معکوس در یک FSM کدگذاری می‌شوند و سپس لاگ‌ها با GFlowNet تولید می‌شوند. در این فرایند، فضای کنش در هر گام زمانی توسط FSM به‌صورت پویا ماسک می‌شود تا لاگ‌ها از نظر ساختاری معتبر تولید شوند. مدل با استفاده از یک هدف Flow Matching و یک پاداش ترکیبی که انطباق با FSM و وفاداری آماری را متوازن می‌کند، آموزش داده می‌شود و در نتیجه داده‌های رفتاری متنوعی مناسب برای طبقه‌بندی قصد تولید می‌کند.

روش‌های ترکیبی

روش‌های ترکیبی، مبنا را از هر دو منبع، یعنی مدل مرجع و لاگ رخداد واقعی استخراج می‌کنند. این رویکردها معانی فرایند یا محدودیت‌های مستخرج از مدل (برای مثال لایه‌های Petri net) را با مؤلفه‌های مبتنی بر یادگیری ماشین تلفیق می‌کنند و با بهره‌گیری از نقاط قوت مکمل دو رویکرد مبتنی بر شبیه‌سازی و مبتنی بر یادگیری ماشین، موجب بهبود واقع‌گرایی، انطباق فرایند، تفسیرپذیری، منطق علّی و انعطاف‌پذیری تجربی می‌شوند.

برخلاف رویکردهای یادگیری عمیق شرطی‌شده با برچسب، روش PNC-SYN از شرطی‌سازی صریح مبتنی بر شبکه پتری استفاده می‌کند تا اعتبار ساختاری را اعمال کند و هم‌زمان از مدل‌های عصبی برای دستیابی به واقع‌گرایی آماری بهره می‌گیرد. PNC-SYN یک LSTM خودرگرسیو چندسر را بر روی لاگ رخداد واقعی آموزش می‌دهد(Kuhn 2026) . به‌طور هم‌زمان، یک سازوکار مبتنی بر شبکه پتری، توزیع خودرگرسیو را تنها به گذارهای فعال شبکه محدود می‌کند تا تضمین شود تمام مسیرهای مصنوعی تولیدشده کامل و بدون خطا هستند. در این فرایند، فعالیت بعدی نمونه‌برداری می‌شود، گذار مربوطه در شبکه پتری اجرا می‌شود تا

[1] Ground-truth plans

نشان‌گذاری محاسبه شود و برچسب‌های زمانی از طریق آنرمال‌سازی و افزودن زمان‌های پیش‌بینی‌شده بین رخدادها بازسازی شوند. اگرچه شرطی‌سازی بر شبکه پتری، سازگاری ساختاری را بهبود می‌دهد اما ممکن است شباهت به توزیع اصلی رخدادها را کاهش دهد.

DeepSimulator. تولید توالی فعالیت‌ها را به یک مدل فرایند تصادفی و مدل‌سازی رفتار زمانی را به شبکه‌های عصبی واگذار می‌کند (Camargo et al. 2022). این روش، مدل فرایند را با استفاده از افزونه Split Minerو trace alignmentدر نرم افزار ProM استخراج می‌کند و برای بهینه‌سازی ابرپارامترهای کشف فرایند، از بهینه‌سازی بیزی و تابع زیان Control-Flow Log Similarity (CFLS)استفاده می‌کند. زمان شروع نمونه فرایند با Prophetپیش‌بینی می‌شوند و دو شبکه LSTM پشته‌ای زمان اجرای فعالیت و زمان انتظار را تولید می‌کنند. این روش همچنین از ویژگی‌های زمانی، کار در حال انجام، اشغال منابع، ورودی‌های n-gram و embeddingهای طبقه‌ای استفاده می‌کند تا پویایی‌های زمانی وابسته به بار کاری را ثبت کند.

مطالعه (Goedertier et al. 2009) کشف فرایند را به‌عنوان یک مسئله طبقه‌بندی دودویی در نظر گرفته است. روش اصلی این مطالعه شامل تکمیل لاگ رخداد با AGNEها است که نشان‌دهنده انتقال‌های حالت غیرمحتمل هستند و فرض می‌شود که رخ نمی‌دهند. برای هر موقعیت رویداد در یک مسیر، با بررسی اینکه آیا توالی مشابهی در میان سایر مسیرهای موجود در لاگ وجود دارد یا خیر، ارزیابی می‌شود که آیا یک فعالیت جایگزین موردنظر می‌توانسته است در آن موقعیت رخ دهد یا خیر. اگر چنین رفتاری یافت نشود، آن رخداد به‌عنوان رخداد منفی علامت‌گذاری می‌شود. این رخدادهای منفیِ مصنوعی، لاگ رخداد مثبت را تکمیل می‌کنند. برای کنترل فرایند استنتاج، این روش از یک پارامتر windowSize به منظور محدود کردن زمینه تاریخی در شرایط وجود حلقه و ساختارهای هم‌زمان، و همچنین از پارامتر negative event injection probabilityبرای جلوگیری از عدم‌توازن کلاس‌ها در مجموعه داده نهایی استفاده می‌کند. پیش‌شرط‌های توالی محلی با استفاده از درخت تصمیمTILDE با روش‌های اکتشافی بهره C4.5 و هرس و مبتنی بر زبان محدودکننده گزاره‌ای no-sequelآموزش داده می‌شوند و در نهایت، قواعد منطقی آموخته‌شده به شبکه‌های پتری گرافیکی تبدیل می‌شوند. شکل ۳ جزییات الگوریتم‌های مورد استفاده برای تولید داده مصنوعی و محیط پیاده‌سازی را نشان می‌دهد. تنها دو مقاله (Broucke et al. 2012 , Goedertier et al. 2009) مبتنی بر جاوا بوده و به توسعه افزونه پروم پرداخته است. سایر ابزارها مبتنی بر پایتون توسعه داده شده‌اند (Camargo et

al. 2022, Redis et al. 2024, Wangelik et al. 2025, Camargo et al. 2021, Graziosi et al. 2025, Singh et al. 2024, Kuhn et al. 2025, and 2026, da Silva Santos et al. 2024, Li et al 2024,) همچنین ۳ مقاله کد منبع باز ارائه نکرده‌اند. یک تحلیل زمانی قابل توجه این است که از سال ۲۰۲۲ ابزارها کاملا به سمت پیاده سازی تحت پایتون حرکت کرده‌اند؛ روندی که تا قبل از سال ۲۰۱۷ اصلا وجود نداشته است.

۲.٤ پرسش دوم: مهم‌ترین کاربردهای لاگ رخداد مصنوعی در مطالعات فرایندکاوی چیست؟

با بررسی دقیق مقالات ارائه شده اهداف توسعه این روش‌ها را می‌توان در چند محور کلیدی شامل ارزیابی الگوریتم‌های فرایندکاوی، حفظ محرمانگی، تولید داده‌های منفی، تولید داده‌های چند بعدی و تحلیل سناریو خلاصه کرد. جدول ۱ هدف اصلی هر مقاله را نشان می‌دهد و نمودار شکل ۴ براساس اهداف چندگانه مقالات که در متن مطالعات ذکر شده بود رسم شده است.

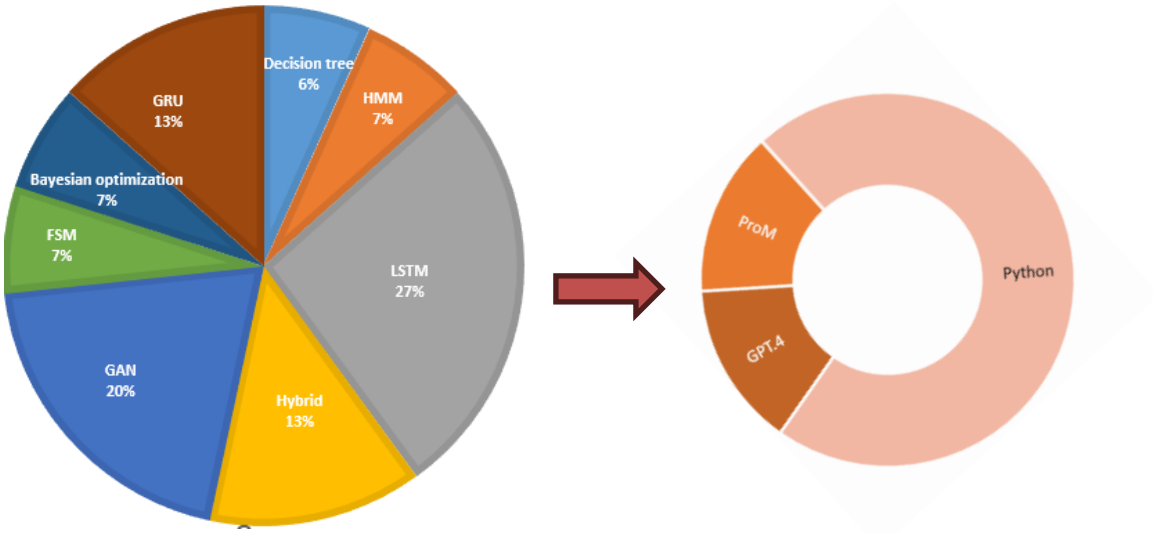
جدول۱- اهداف توسعه روش تولید لاگ رخداد مصنوعی

مقالات	اهداف توسعه روش تولید لاگ رخداد مصنوعی
(Camargo et al. 2021), (Redis et al. 2024) (Maldonado et al. 2024), (Singh et al. 2024), (Samanta 2025)	ارزیابی الگوریتم‌های فرایندکاوی
(Li et al. 2024), (Kuhn et al. 2026), (Yang et al. 2017), (Wangelik et al. 2025) ,(Kuhn et al. 2025)	حفظ محرمانگی
(Broucke et al. 2012), (Goedertier et al. 2009)	تولید داده های منفی
(da Silva Santos et al. 2024), (Graziosi et al. 2025)	تولید داده چندبعدی
(Camargo et al. 2022)	تحلیل سناریو

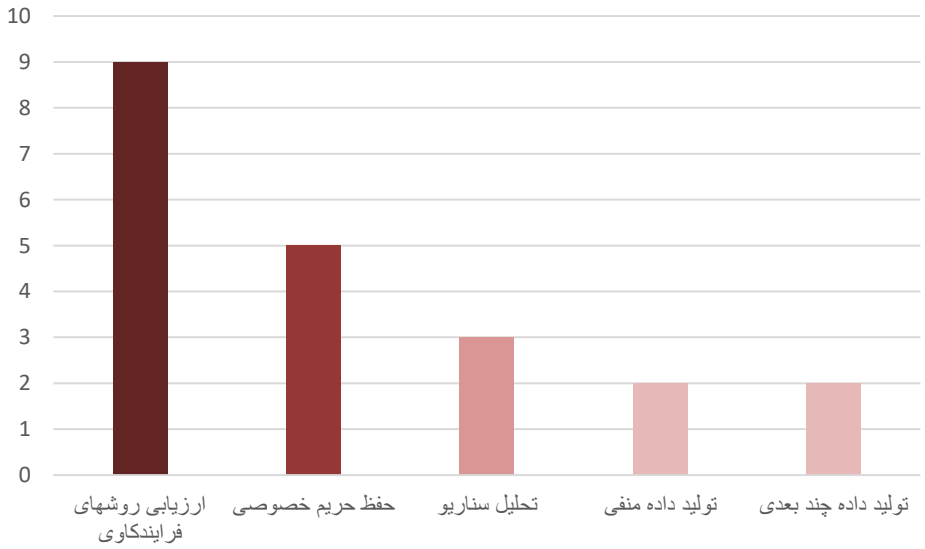
۳.٤ پرسش سوم: مطالعات مختلف کیفیت، سودمندی و کارایی لاگ رخداد مصنوعی را با استفاده از چه معیارهایی ارزیابی کرده‌اند؟

بسته به نوع الگوریتم پیاده‌سازی شده، هر روش از معیارهای خاصی برای ارزیابی و سنجش عملکرد استفاده کرده است. فقدان چارچوب و معیارهای ارزیابی واحد برای مقایسه ابزارها و روش‌های توسعه داده شده در مطالعات به چشم می‌خورد. این مسئله چالشی برای مقایسه روش

پیشنهادی با ابزارها و روش‌های پیشین است. به‌طور کلی ارزیابی روش‌ها و ابزارهای توسعه داده شده برای تولید لاگ رخداد مصنوعی اغلب در دو گام انجام شده است.



شکل۳- الگوریتم‌های مورد استفاده و محیط‌های پیاده‌سازی



شکل٤- دسته بندی مقالات براساس اهداف تولید لاگ رخداد مصنوعی

۱-ارزیابی براساس شباهت: در این گام به سنجش شباهت و بررسی میزان تطابق آماری داده‌های مصنوعی با داده‌های واقعی پرداخته می‌شود و فاصله بین ویژگی‌های هدف و ویژگی‌های لاگ مصنوعی تولیدشده اندازه‌گیری می‌شود(Graziosi et al. 2025, Wangelik et al. 2025,Kuhn et al. 2025, Kuhn et al. 2026 ,Maldonado et al. 2024, Li et al. 2024, Camargo et al. 2019,Camargo et al. 2022, Camargo et al. 2021). ازجمله معیارهای مورد استفاده عبارتند از: معکوس فاصله[1]، فاصله انتقال زمین[۲]، فاصله کولموگروف–اسمیرنوف[۳]، فاصله هلینگر[۴]، فاصله لونشتاین، محاسبه تفاوت مطلق بین لاگ واقعی و مصنوعی، فاصله لاگ کنترل- جریان[۵] ، فاصلهn-gram ،فاصله Dynamic Time Warping، محاسبه پوشش فضای ویژگی‌ها توسط داده‌های تولیدشده با تحلیل PCAو Convex Hull، واگرایی کولبک–لایبلر[۶] ، فاصله کای‌دو[۷]، آنتروپی و میزان هم‌پوشانی بیگرام‌ها[8].

۲- ارزیابی براساس کاربردپذیری: مدل‌های فرایند داده‌های مصنوعی تولید شده به توسط الگوریتم‌های کشف مختلف مانند آلفا، ژنتیک ماینر، ایناداکتیو ماینر و هیوریستیک ماینر استخراج شده و کاربردپذیری لاگ‌های تولیدشده در فرایندکاوی ارزیابی می‌شود(Goedertier et al. 2009, Yang et al. 2017, Li et al. 2024, Singh et al. 2024, Wangelik et al. 2025, Graziosi et al. 2025, Kuhn et al. 2025, Maldonado et al. 2024, Samanta, 2025). مدل فرایند با معیارهای رایج تحلیل فرایند مانند برازش[۹]، طول مسیرها، دقت[۱۰]، واریانس توالی[۱۱]، اندازه مدل براساس تعداد گره‌ها و

---

[1] حداقل تعداد عملیات ویرایشی (درج، حذف، جایگزینی و جابجایی) برای تبدیل یک رشته (مسیر) به رشته دیگر.

[۲] Earth-Mover's Distance : حداقل هزینه برای تبدیل یک توزیع (هیستوگرام) به توزیع دیگر. این معیار برای مقایسه زمان‌های پردازش یا توزیع واریانت‌ها استفاده می‌شود.

[3] Kolmogorov–Smirnov

[4] Hellinger Distance

[5] Control-Flow Log Distance

[6] Kullback–Leibler Divergence

[7] Chi-squared Distance

[8] Bigram Overlap

[۹] Fitness : با Replay کردن مسیرها روی شبکه پتری محاسبه می‌شود و جریمه‌هایی برای توکن‌های گم‌شده و توکن‌های باقی‌مانده در نظر می‌گیرد.

[۱۰] Accuracy : میانگین وزنی نرخ منفی واقعی و نرخ مثبت واقعی. برای حذف سوگیری ناشی از تعداد نابرابر نمونه‌های مثبت و منفی، وزن برابر به هر دو داده می‌شود.

[11] Sequence Variance

پیچیدگی مدل[۱] ارزیابی شده‌اند. سایر معیارهای مورد استفاده عبارتند از: بازخوانی رفتاری[۲]، ویژگی رفتاری[۳] ، نرخ مثبت و منفی واقعی[۴]، معیار تجزیه[۵] ، مناسب‌بودن رفتاری پیشرفته[6]، میزان تفاوت بین توزیع فعالیت‌ها در داده‌های مصنوعی و دقت طبقه‌بندی خبرگان، خطای وقوع نوع فعالیت[۷]، خطای زمانی فعالیت‌ها[8]، فراخوانی، F1-score، حفظ ساختار داده واقعی[۹] ، تنوع[10]، نوآوری[۱۱] ، بررسی توزیع وقوع انواع فعالیت‌ها[۱۲] ، هزینه‌های هم‌ترازی [۱۳] ، توزیع زمان چرخه فرآیند[۱۴] ، زمان کل فرایند[۱۵] ، توزیع نسبی رخدادها[۱۶] و طبقه‌بندی نیت[۱۷] برای تعمیم مدل‌های آموزش‌دیده با داده‌های مصنوعی بر روی داده‌های واقعی.

مطالعه (Meneghello et al. 2023) معیارهای عددی دقیقی برای ارزیابی گزارش نکرده است و تنها به مقایسه با روش‌های سنتی شبیه‌سازی رویدادگسسته و Dsim پرداخته است. در مطالعه (Redis et al. 2024) برای مقایسه عملکرد پرس‌وجوهای اصلی و نسخه‌های بازنویسی‌شده [18] از آزمون Wilcoxon Signed-Rankاستفاده شده است. معیارهای ارزیابی شامل برازش بازپخش[۱۹]، برازش هم‌ترازی[۲۰]، طول برنامه[21]، پیچیدگی کاردوسو[۱]، پیچیدگی سیکلوماتیک و

---

[1] Control-Flow Complexity

[۲] Behavioral Recall :همان نرخ مثبت واقعی است.

[۳] Behavioral Specificity : یا نرخ منفی واقعی، درصد رخدادهای منفی مصنوعی که مدل به درستی آن‌ها را غیرمجاز تشخیص داده است.

[۴] True Positive / Negative Rate : به معنی درصد رخدادهای مثبت / منفی که در لاگ وجود دارند و توسط مدل به درستی شناسایی می‌شوند

[۵] Parsing Measure : نسبت تعداد مسیرهایی از لاگ که به طور کامل توسط مدل فرآیندی قابل پیمایش هستند به کل مسیرها.

[6] Advanced Behavioral Appropriateness : سنجش میزان دقت رفتاری بر اساس روابط تقدم و تاخر که مدل اجازه می‌دهد، در مقایسه با آنچه در لاگ مشاهده شده است.

[7] Activity Type Occurrence Error
[8] Activity Timestamp Error
[9] Fidelity
[10] Diversity
[11] Novelty
[12] Activity Type Occurrence Distribution
[13] Alignment Costs
[14] Cycle Time Distribution
[15] Throughput Time
[16] Relative Event Distribution
[17] Intent Classification
[18] Paraphrased Queries
[19] Replay Fitness
[20] Alignment Fitness
[21] Plan Length

درجه هم‌زمانی[۲] بوده‌اند. اعتبارسنجی مدل با مقایسه برنامه‌های تولیدشده توسط مدل زبانی با مدل‌های مرجع انجام شده است به این صورت که مدل‌های مرجع ابتدا به قالب پتری تبدیل شده و سپس میزان انطباق برنامه‌های تولیدشده با آن‌ها بررسی شده است. مجموعه‌داده مورد استفاده شامل ۵۳۲ پرس‌وجو از زیرمجموعه TaskBench Multimedia بوده که هر پرس‌وجو بین ۵ تا ۶ بار بازنویسی شده و در مجموع حدود ۱۹۶۵ نمونه برای تحلیل هم‌ترازی ایجاد شده است. Terpsichora با معیارهایی شامل اندازه مدل، تراکم[۳]، قابلیت تفکیک‌پذیری[۴]و تنوع محدودیت‌ها[5] ارزیابی شده است(da Silva Santos et al. 2024). اعتبارسنجی روش از طریق تولید لاگ‌های رخداد از مدل‌های فرایندی مصنوعی و سپس بازکشف مدل‌ها از روی این لاگ‌ها انجام شده است. مجموعه‌داده مورد استفاده شامل ۵۰۰ مدل فرایندی تولیدشده مصنوعی برای حوزه‌های کسب‌وکار مختلف بوده است.

داده‌های ارزیابی شامل لاگ‌های رخداد عمومی هستند که ازجمله آن‌ها می‌توان به لاگ رخداد *Helpdesk* (Singh et al. 2024,Camargo et al. 2019) ، لاگ رخداد *Sepsis* (Maldonado et al. 2024, Li et al. 2024, Graziosi et al. 2025, Wangelik et al. 2025, Kuhn et al. 2026, Kuhn et al. 2025) لاگ رخداد *BPIC-2012* (Singh et al. 2024, Maldonado et al. 2024, Camargo et al. 2019, Graziosi et al. 2025, 2021,2022, Li et al. 2024, Wangelik et al. 2025)، لاگ رخداد *BPIC-2013* (Maldonado et al. 2024, Camargo et al. 2019, Wangelik et al. 2025)، لاگ رخداد *BPIC-2015,2020* (Maldonado et al. 2024)، لاگ رخداد *Traffic_Fines* (Graziosi et al. 2025, Kuhn et al. 2025, Maldonado et al. 2024) و لاگ رخداد *Hospital billing* (Kuhn et al. 2025). برخی مطالعات از لاگ رخداد مصنوعی و لاگ های رخداد مطالعه موردی برای ارزیابی استفاده کرده‌اند(Camargo et al. 2021,2022 ,Yang et al. 2017,Goedertier et al. 2009, Li et al. 2024, Goedertier et al. 2009, Samanta, 2025,).

---

[۱] Cardoso and Cyclomatic Complexity : معیارهایی برای سنجش پیچیدگی ساختاری مدل‌های کشف شده

[۲] Degree of Concurrency : به صورت نسبت بین طولانی‌ترین و کوتاه‌ترین مسیر در یک شبکه پتری تعریف شده است.

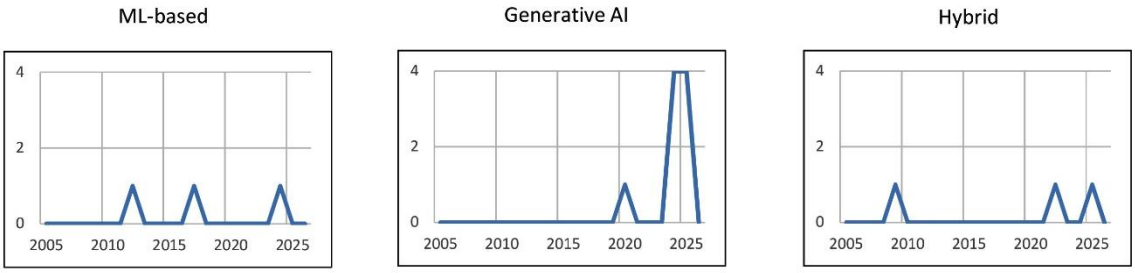
[3] Density

[4] Separability

[5] Constraint Variability

۵. بحث و نتیجه گیری

کمبود داده، محرمانگی اطلاعات و محدودیت‌های حریم خصوصی دسترسی به لاگ‌های رخداد واقعی را دشوار کرده و توسعه و ارزیابی روش‌های فرآیندکاوی را محدود ساخته است. بنابراین تولید داده مصنوعی راهکاری مؤثر محسوب می‌شود. این مطالعه، گزارشی از یک مرور نظام‌مند برروی روش‌های تولید داده مصنوعی مبتنی بر یادگیری ماشین و یادگیری عمیق در فرایندکاوی می‌دهد. بررسی سیستماتیک ۱۵ مطالعه پیشگام منتشر شده تا آگوست ۲۰۲۶، حول چهار سوال تحقیقاتی اصلی ساختار یافته است تا دیدگاهی چندگانه در مورد چشم‌انداز تحقیق ارائه دهد. شکل ۵ روند زمانی چاپ مقالات را نشان میدهد. همانطور که مشاهده میشود روشهای مبتنی بر یادگیری عمیق و هوش مولد از سال ۲۰۲۳ مورد توجه فزاینده توسط محققان قرار گرفته اند. درحالی‌که به روش‌های دیگر مانند روش‌های ترکیبی کمتر پرداخته شده است.



شکل۵- روند زمانی چاپ مقالات مورد بررسی

بررسی مطالعات نشان‌دهنده تغییر تدریجی به سمت یک الگوی یادگیری ماشین متنوع‌تر و داده‌محور است. رویکردهای سنتی مبتنی بر شبیه‌سازی قوانین و الزامات یا محدودیت‌های کنترل-جریان را در خود جای می‌دهند و بنابراین به صحت، سلامت و قابلیت حسابرسی ساختاری برتر دست می‌یابند. محدودیت اصلی این روش‌ها، پیچیدگی مهندسی بالای مدل‌های رویه‌ای، اسکریپت‌نویسی دستی و فرضیات بیش از حد ساده‌شده هنگام مواجهه با گزارش‌های رویداد بدون ساختار است که اغلب منجر به توزیع تأخیرهای غیرواقعی و مستقل از صف می‌شود. روش‌های مبتنی بر یادگیری عمیق و یادگیری ماشین از جمله GANها، خودرمزگذارها، مدل‌های انتشار، خودرگرسیوها و مدل‌های زبانی به عنوان رویکردهای نویدبخش برای تولید لاگ رخداد مصنوعی ظهور کرده‌اند. GANها به ویژه برای لاگ‌های رخداد پیچیده و چند بعدی با مسیرهای موازی متغیر و وابستگی‌های دوربرد به خوبی عمل می‌کنند. آن‌ها داده‌های مصنوعی با دقت بالا

تولید می‌کنند و در جایی‌که حریم خصوصی مهم تلقی می‌شود برجسته باقی می‌مانند. با این حال، مقداردهی‌های اولیه ، ابرپارامترها و توقف زودهنگام ممکن است باعث بی‌ثباتی آموزش، فروپاشی حالت یا انطباق رفتاری متفاوت شود. شبکه‌های عصبی عموما هزینه‌های محاسباتی بالاتری نسبت به روش‌های مبتنی بر شبیه‌سازی دارند. بنابراین GANها زمانی مناسب هستند که واقع‌گرایی رفتاری و تقویت وظیفه‌محور در اولویت باشند و منابع محاسباتی امکان تنظیم گسترده را فراهم کنند. معماری‌های AE و CVAE آموزش پایدار، فضاهای نهفته تفسیرپذیر و ادغام ویژگی‌های مسیر و الگوهای رخداد گمشده را ارائه می‌دهند که به تولید لاگ رخداد کامل و توالی‌های خلاف واقع منجر می‌شود. چالش‌های این روش‌ها شامل هموارسازی بیش از حد ویژگی‌های زمانی پیوسته شبیه گاوسی، عدم نمایش متغیرهای نادر و وفاداری داده‌ها در حدی پایین‌تر نسبت به GANها است. بنابراین VAEها و CVAEها زمانی‌که آموزش قوی، عوامل پنهان ساختاریافته و تحلیل سناریوی قابل کنترل بر وفاداری اولویت دارند روش مناسبی هستند. DDPMها در برابر فروپاشی حالت مقاوم هستند و اغلب مسیرهایی با وفاداری بالا تولید می‌کنند، اما نویززدایی تکراری باعث هزینه‌های محاسباتی قابل توجهی می‌شود. همچنین قابلیت تفسیرپذیری محدودی دارند و تضمین‌های حریم خصوصی عمدتا تجربی هستند. این روش‌ها زمانی ترجیح داده می‌شوند که تنوع مسیر و وفاداری به داده‌ها، محاسبات قابل توجه را توجیه کنند. مدل‌های خودرگرسیو ضمن پشتیبانی از شرطی سازی دقیق پارامترهای نمونه، انسجام کنترل–جریان و زمان را حفظ می‌کنند. چالش‌ها شامل سوگیری نوردهی، هزینه‌های بالای راه‌اندازی و فواصل زمانی نامنظم است. این روش‌ها برای اجرای دقیق نمونه‌ها و تعیین گلوگاه‌ها بهترین هستند. مدل‌های زبانی بزرگ علیرغم مزایایشان، مستعد توهم، رانش منطقی و نقض محدودیت‌های فرآیند هستند و خروجی‌هایی تولید می‌کنند که ممکن است از نظر آماری، معنایی یا ساختاری با رفتار فرآیند واقعی همسو نباشند. پنجره‌های زمینه‌ی محدود، تولید فرآیند در مقیاس بزرگ را محدود می‌کنند. پارامترها همچنین نیاز به ایجاد تعادل بین تنوع و تکرارپذیری دارند، زیرا دماهای بالاتر تنوع را افزایش می‌دهند اما تکرارپذیری را کاهش می‌دهند.

در مقایسه با روش‌های سنتی مبتنی بر شبیه‌سازی، رویکردهای داده‌محور نیاز به زمان‌های آموزش طولانی و پیکربندی‌های سخت‌افزاری با کارایی بالا (به عنوان مثال GPUها) دارند. فرآیند تولید داده در این روش‌ها نیز به دلیل نمونه‌برداری توالی بازگشتی یا زنجیره‌های مارکوف معکوس

طولانی کندتر است. با این حال، اگر به درستی آموزش داده شوند، این مدل‌ها داده‌هایی با کیفیت آماری، رفتاری و زمانی بهتر تولید می‌کنند. تحقیقات آینده باید ترکیب متامدل‌های رسمی (به عنوان مثال MP-Declare) را با مدل‌های بزرگ زبانی را بررسی کنند. این امر به تکنیک‌های پیشرفته مهندسی دستور امکان می‌دهد تا ساختارهای مسیر بسیار تکرارپذیر و منطبق را تولید کنند. زیرا مدل‌های زبانی به تنهایی مدل‌هایی احتمالاتی هستند و فاقد سازوکارهای ذاتی برای رعایت قوانین فرآیند می‌باشند بنابراین ترکیب آن‌ها با مدل‌های پتری اجازه می‌دهد که محدودیت‌های ساختاری فرآیند اعمال شده و از تولید رفتارهای نامعتبر جلوگیری شود.

تشکر و قدردانی

این پژوهش با حمایت مالی بنیاد ملی نخبگان و در قالب تعهدات طرح پسادکتری شهید چمران انجام شده است.

فهرست منابع

1- Yari Eili M., Rezaeenour, J., Jalaly, B. A., & Bijani, S. (2023). Analyzing the research grant process in Iran's National Elites Foundation: An approach based on process mining and machine learning

2- G. Acampora, A. Vitiello, B. Di Stefano, W. M. P. van der Aalst, C. W. Günther, and H. M. W. Verbeek, "IEEE 1849™: The XES Standard: The Second IEEE Standard Sponsored by IEEE Computational Intelligence Society," *IEEE Computational Intelligence Magazine*, pp. 4-8, 2017.

3- Kamal, O. S., Sohail, S. A., & Bukhsh, F. A. (2024). Optimizing privacy-utility trade-off in healthcare processes: Simulation, anonymization, and evaluation (using process mining) of event logs. In *14th International Conference on Simulation and Modeling Methodologies, Technologies and Applications SIMULTECH 2024* (pp. 289-296). SCITEPRESS.

4- Yari Eili, M., Vafadar, S., Rezaeenour, J., & Sharif-Alhoseini, M. (2022). Self-service registry log builder: a case study in national trauma registry of Iran. *Methods of Information in Medicine*, *61*(05/06), 185-194.

5- Salehi, A., Aghdasi, M., Khatibi, T., & Sheikhmohammady, M. (2023). A Conceptual Framework for Preprocessing and Improving Quality of Event Log in Process Mining. *Iranian Journal of Information Processing and Management*, *38*(3), 945-979.

6- Andrea, B., Re, B., Rossi, L., & Tiezzi, F. (2025). A framework for purpose-guided event logs generation. *Data & Knowledge Engineering*, *161*.

7- Di Ciccio, C., Bernardi, M. L., Cimitile, M., & Maggi, F. M. (2015). Generating event logs through the simulation of declare models. In *Workshop on Enterprise and Organizational Modeling and Simulation* (pp. 20-36).

8- Skydanienko, V., Di Francescomarino, C., Ghidini, C., & Maggi, F. M. (2018). A tool for generating event logs from multi-perspective declare models. *Proceedings of the Dissertation Award, Demonstration, and Industrial Track at BPM 2018*, *2196*, 111-115.

9- Jouck, T., & Depaire, B. (2019). Generating artificial data for empirical analysis of control-flow discovery algorithms: A process tree and log generator. *Business & Information Systems Engineering*, *61*(6), 695-712.

10- Kuhn, M., Grüger, J., Matheja, C., & Rivkin, A. (2024, September). Data petri nets meet probabilistic programming. In *International Conference on Business Process Management* (pp. 21-38).

11- Nedopetalski, F., & de Freitas, J. C. J. (2021, April). Process mining and simulation for a p-time petri net model with hybrid resources. In *2021 Systems and Information Engineering Design Symposium* (pp. 1-6).

12- Friederich, J., Khodadadi, A., & Lazarova-Molnar, S. (2026). PySPN: a Python library for stochastic Petri net modeling, simulation, and event log generation. *Simulation*, *102*(1), 59-73.

13- Fonseca, J., & Bacao, F. (2023). Tabular and latent space synthetic data generation: a literature review. *Journal of Big Data*, *10*(1), 115.

14- Lautrup, A. D., Hyrup, T., Zimek, A., & Schneider-Kamp, P. (2024). Systematic review of generative modelling tools and utility metrics for fully synthetic tabular data. *ACM Computing Surveys*, *57*(4), 1-38

15- Goyal, M., & Mahmoud, Q. H. (2024). A systematic review of synthetic data generation techniques using generative AI. *Electronics*, *13*(17), 3509.

16- Malin, B., & Goodman, K. (2018). Between access and privacy: challenges in sharing health data. *Yearbook of medical informatics*, *27*(01), 055-059.

17- Fahrenkrog-Petersen, S. A., van der Aa, H., & Weidlich, M. (2020, September). PRIPEL: privacy-preserving event log publishing including contextual information. In *International Conference on Business Process Management* (pp. 111-128). Cham: Springer International Publishing.

18- A. Pawar, S. Ahirrao, P.P. Churi, Anonymization techniques for protecting privacy: A survey, in: 2018 IEEE Punecon, 2018, pp. 1–6,

19- Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. Journal of Official Statistics 9, 462–468.

20- Bellovin, S. M., Dutta, P. K., & Reitinger, N. (2019). Privacy and synthetic datasets. *Stan. Tech. L. Rev.*, *22*, 1.

21- James, S., Harbron, C., Branson, J., & Sundler, M. (2021). Synthetic data use: exploring use cases to optimise data utility. *Discover Artificial Intelligence*, *1*(1), 15.

22- De Medeiros, A. A., & Günther, C. W. (2005). Process mining: Using CPN tools to create test logs for mining algorithms. In *6th Workshop and Tutorial on Practical Use of Coloured Petri Nets and the CPN Tools (CPN'05), October 24-26, 2005, Aarhus, Denmark* (pp. 177-190). University of Aarhus.

23- Xiao, Y., & Watson, M. (2019). Guidance on conducting a systematic literature review. *Journal of planning education and research*, *39*(1), 93-112.

24- Vanden Broucke, S., Weerdt, J.D., Baesens, B., Vanthienen, J.: Improved artificial negative event generation to enhance process event logs. Advanced Information Systems Engineering- 24th International Conference, CAiSE 2012, Lecture Notes in Computer Science, vol. 7328, pp. 254–269. 63

25- Yang, S., Zhou, Y., Guo, Y., Farneth, R. A., Marsic, I., & Burd, R. S. (2017, August). Semi-synthetic trauma resuscitation process data generator. In *Proceedings. IEEE International Conference on Healthcare Informatics* (Vol. 2017, p. 573).

26- Maldonado, A., Frey, C. M., Tavares, G. M., Rehwald, N., & Seidl, T. (2024, September). GEDI: Generating event data with intentional features for benchmarking process mining. In *International Conference on Business Process Management* (pp. 221-237). Cham: Springer Nature Switzerland.

27- Bauer, A., Trapp, S., Stenger, M., Leppich, R., Kounev, S., Leznik, M., ... & Foster, I. (2024). Comprehensive exploration of synthetic data generation: A survey. *arXiv preprint arXiv:2401.02524*.

28- Lautrup, A. D., Hyrup, T., Zimek, A., & Schneider-Kamp, P. (2024). Systematic review of generative modelling tools and utility metrics for fully synthetic tabular data. *ACM Computing Surveys*, *57*(4), 1-38.

29- Camargo, M., Dumas, M., & González-Rojas, O. (2021). Discovering generative models from event logs: data-driven simulation vs deep learning. *PeerJ Computer Science*, *7*, e577.

30- Kuhn, M., Grüger, J., & Bergmann, R. (2025). PALSYN: a method for synthetic multi-perspective event log generation with differential private guarantees. *Process Science*, *2*(1), 28.

31- Graziosi, R., Ronzani, M., Buliga, A., Di Francescomarino, C., Folino, F., Ghidini, C., ... & Pontieri, L. (2025). Generating multiperspective process traces using conditional variational autoencoders. *Process Science*, *2*(1), 8.

32- Li, K., Yang, S., Sullivan, T. M., Burd, R. S., & Marsic, I. (2024). ProcessGAN: Generating privacy-preserving time-aware process data with conditional generative adversarial nets. *ACM transactions on knowledge discovery from data*, *18*(9), 1-31.

33- Wangelik, F., Rafiei, M., Pourbafrani, M., & van der Aalst, W. M. (2025). Releasing differentially private event logs using generative models. *Data & Knowledge Engineering*, *159*, 102450.

34- Singh, A., Bettouche, Z., & Fischer, A. (2024, September). Synthetic training-data generation for ml-based process mining tools. In *2024 14th International Conference on Advanced Computer Information Technologies (ACIT)* (pp. 705-709). IEEE

35- Long, L., Wang, R., Xiao, R., Zhao, J., Ding, X., Chen, G., & Wang, H. (2024, August). On LLMs-driven synthetic data generation, curation, and evaluation: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 11065-11082).
36- da Silva Santos, W., Coutinho, J. R., Baião, F., Spyrides, G. M., & Lopes, H. C. V. (2024, October). Terpsichora: a tool to generate fsm: Springer Nature Switzerland.
37- Redis, A. C., Sani, M. F., Zarrin, B., & Burattin, A. (2024). Processtbench: an LLM plan generation dataset for process mining. *arXiv preprint arXiv:2409.09191*.
38- Samanta, R. (2025). Structurally Valid Log Generation using FSM-GFlowNets. *arXiv preprint arXiv:2510.26197*.
39- Kuhn, M., Trinh, T., Grüger, J., & Bergmann, R. (2026). Synthetic Log Generation Under Control-Flow Conditions Using Autoregressive Models. In *International Conference on Advanced Information Systems Engineering* (pp. 12-24). Springer, Cham.
40- Camargo, M., Dumas, M., & González-Rojas, O. (2022, June). Learning accurate business process simulation models from event logs via automated process discovery and deep learning. In *International Conference on Advanced Information Systems Engineering* (pp. 55-71). Cham: Springer International Publishing.
41- Goedertier, S., Martens, D., Vanthienen, J., & Baesens, B. (2009). Robust Process Discovery with Artificial Negative Events. *Journal of Machine Learning Research*, *10*(6). 80

A survey on machine learning-driven synthetic event log generation

Jalal Rezaeenour*, Mansoureh Yari Eili
PhD in Industrial Engineering, Professor, Department of Industrial Engineering, University of Qom, Qom, Iran: J.rezaee@qom.ac.ir

PhD of Information Technology Engineering, Post Doc Researcher, Department of Industrial Engineering, University of Qom, Qom, IranM.yari@stu.qom.ac.ir

Abstract
The scarcity of real-world data, driven by privacy constraints, data collection costs, and poor data quality, poses a major challenge to the development and evaluation of process mining algorithms. Synthetic event logs that preserve the statistical and structural characteristics of real processes while maintaining no direct linkage to real-world instances offer an effective means of addressing these limitations. This study presents a systematic review of machine learning–based methods for synthetic event log generation in process mining. A systematic search was conducted in ScienceDirect, PubMed, IEEE Xplore, and SpringerLink through August 2025. Of 1,990 initially identified records, 15 studies were ultimately included following

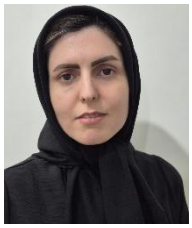
duplicate removal and a three-stage screening process (title, title and abstract, and full text).
Machine learning has emerged as an advanced paradigm for synthetic event log generation, encompassing an evolutionary spectrum from traditional machine learning approaches, such as decision trees, to advanced deep learning and generative artificial intelligence methods, including language models and autoencoders, as well as hybrid approaches. The reviewed studies indicate a growing emphasis on deep learning- and generative AI–based methods. The primary objectives of synthetic event log generation were algorithm evaluation (5 studies, 33%), confidentiality preservation (5, 33%), negative event generation (2, 13%), multidimensional data generation (2, 13%), and scenario analysis (1, 7%). Fidelity, defined as the statistical similarity between synthetic and real data, was assessed in all studies. However, process mining utility, defined as the evaluation of process discovery or conformance checking using synthetic data, was reported in only eight studies (53%), while privacy was explicitly evaluated in only two (13%). The absence of an integrated evaluation framework and the limited assessment of privacy remain major challenges for the reliable comparison of existing approaches.
Future research should focus on integrating the generative capabilities of advanced models with structured process models, such as Petri nets, to ensure the structural validity of generated logs through the integration of data and domain knowledge. Comparative evaluation of different generation mechanisms is also needed to determine how the choice of generation mechanism affects the performance of downstream process mining algorithms. A fundamental question therefore remains unresolved: which generation mechanism(s) can produce synthetic event logs that are most effective for process mining?

Keywords: Synthetic Event Log, Process Mining, Machine Learning, Deep Learning, Survey

جلال رضایی نور

متولد سال ۱۳۵۶ دارای مدرک تحصیلی دکتری در رشته مهندسی صنایع از دانشگاه علم و صنعت ایران است و دوره فرصت مطالعاتی خود را در دانشگاه مونش استرالیا گذرانده است. ایشان هم‌اکنون استاد گروه مهندسی صنایع دانشگاه قم است. تحول دیجیتال، مدیریت و مهندسی دانش و مدیریت فرآیندهای کسب و کار و فرآیندکاوی از جمله علایق پژوهشی وی است.

منصوره یاری ایلی



در حال حاضر در حال گذراندن دوره پسادکتری مشترک بین بنیاد ملی نخبگان و گروه مهندسی صنایع در دانشگاه قم است. توسعه الگوریتم‌های مبتنی بر یادگیری ماشین در حوزه‌های فرایندکاوی و به‌طور خاص سلامت و پزشکی از جمله علایق پژوهشی وی است.