Attention Mechanisms in Process Mining: A Systematic Literature Review

Conference Paper · October 2023

DOI: 10.1109/CLEI60451.2023.10346135

3 authors:

Gonzalo Rivera Lazo
Universidad Técnica Federico Santa María
2 PUBLICATIONS **1** CITATION
SEE PROFILE

Hernán Astudillo
Andrés Bello University
230 PUBLICATIONS **2,086** CITATIONS
SEE PROFILE

Ricardo Nanculef
University of Bristol
50 PUBLICATIONS **242** CITATIONS
SEE PROFILE

Attention mechanisms in Process Mining: A Systematic Literature Review

Gonzalo Rivera-Lazo
Departamento de Informática
U. Técnica Federico Santa María
Valparaíso, Chile
0000-0002-3483-0710

Hernán Astudillo
ITiSB - Universidad Andrés Bello
Viña del Mar, Chile
&
Universidad Técnica Federico Santa María
Santiago, Chile
0000-0002-6487-5813

Ricardo Ñanculef
Departamento de Informática
U. Técnica Federico Santa María
Valparaíso, Chile
0000-0003-3374-0198

***Abstract*—Process Mining (PM) focuses on monitoring and optimizing long-running business processes by examining their execution event logs (usually complex and heterogeneous) to obtain insights and enable data-driven decisions. Several Machine Learning (ML) techniques have been recently proposed to exploit these logs as learning datasets and enable examination of past events and predict future ones, but their black-box nature makes hard for human analysts to *interpret* their results and recognize the key parts of input data. *Attention mechanisms* (AM) is an ML technique that does address these shortcomings, but it has been little used for PM. This article describes the design, results and findings of a systematic literature review of attention mechanisms for PM. We addressed three research questions: (a) for which applications are AM used? (b) which kinds of AM are used? and (c) how are AM combined with other ML techniques? An initial search yield 73 papers, and inclusion/exclusion criteria left sixteen, published between 2017 and 2023. Key finding are that: (1) the most common application is *sequence prediction*, (2) most studies combine *global and item-wise attention*, added as layers after an encoder generates the continuous representation, and (3) emerging research topics include *anomaly detection* and *data representation*. This study shows that attention mechanisms can help process analysts to get some sense of *interpretability*, and showcases the bright potential for process mining of attention mechanisms, which paradoxically have received little attention themselves.**

***Index Terms*—Attention mechanisms, Deep Learning, Process Mining, Systematic Literature Review, Interpretability**

I. Introduction

Digital transformation is picking up speed in many industries and organizations [1], Information system handle both the management and operation of a company, and allow to enhance business processes by replacing manual work with (partial or near-total) automation. In addition, it creates complex event logs that describe the actual execution of processes. Process Mining [2] addresses the use of these event logs to extract insights on bottlenecks, conformance checking, decision points, etc. As operational predictive support benefits from process mining techniques, machine learning and statistical methods have been proposed to predict valuable indicators as: the remaining sequence of an ongoing trace, the remaining trace time, etc. Further, deep learning methods like recurrent neural networks [3] have shown results that outperform traditional machine learning, and approaches that integrate attention mechanisms allow models to selectively focus on key parts of the input data. Moreover, such mechanisms strive to eliminate the black-box feature by providing transparency to the internal model behavior associated with each prediction.

Several surveys have explored the use of *attention mechanisms* in other areas, like image processing [4], voice recognition [5], and image captioning [6], but there are very few on their use for Process Mining (only two [3] [7] but omit several important details of the characteristics and applications of the technique), although Neu et al. [3] proposed to continue a research line in *causality and explainability of prediction* on techniques that allow users to understand the model behavior and the logic behind its predictions.

This study presents a systematic literature review on the use of attention mechanisms for Process Mining. Its main contributions are:

- Create a unified presentation of the current literature approaches.
- Classify the selected works by their main characteristics and contributions.
- Suggest promising research opportunities by identifying conflicting declarations and research gaps.

The remainder of the paper is structured as follows: Sections II and III cover process mining and attention mechanisms; Section IV surveys previous related work; Section V describes the SLR methodology; Section VI presents results and findings; and Section VII summarizes and concludes.

II. Business Process Mining

A *business process* [8] is a set of coordinated activities in a technical/organizational environment, which are executed by one or more resources at a given time, and can be related internally and with external entities. Besides, processes have restrictions, both legal regulations and the organization's restrictions. Traditionally Business Process

979-8-3503-1887-6/23/$31.00 ©2023 IEEE

Management (BPM) focuses on modeling and analyzing ideal processes that a business would like to have and to which it hopes to converge [8], but analysis of actual event logs shows [9] that processes are complex, incomplete, and heterogeneous, making hard their analysis and monitoring.

At the intersection of BPM and Data Mining there is Process Mining [9], and aims to extract insights from actual event logs to enable processcentric and fact-based decision-making. Process Mining has [2] [10] three main activities:

- *Process Discovery* entails constructing a model solely from an event log without using any a-priori information. It is often surprising to organizations that existing techniques are capable of discovering real processes based solely on example executions recorded in event logs.
- *Process Conformance checking* involves the comparison of an existing process model with an event log of the same process. This activity is useful in determining whether reality, as recorded in the log, is consistent with the model and vice versa. It is worth noting that various types of models, such as procedural models, organizational models, declarative process models, business rules/policies, laws, etc. can be subjected to conformance checking.
- *Process Enhancement* has the objective is to expand or refine an existing process model by utilizing information obtained from the recorded event log. Unlike conformance checking which measures the degree of agreement between the model and reality, this third type of process mining aims to modify or supplement the a-priori model. For example, utilizing timestamps in the event log can facilitate the expansion of the model to display bottlenecks, service levels, throughput times, and frequencies.

These activities can be addressed from several perspectives regarding analysis, being the main ones Control Flow (focused on *What* activities are happening and how they are related within the process), Organizational (focused on *Who* executes the activity and how this is related to other resources within the process, e.g. equipment, facilities etc.), and Time (focused on time-related issues, like case duration, seasonal behaviors, e.g. When). and *Time* (with focus on time-related issues, like case duration, seasonal behaviors, e.g. *When*).

The life-cycle model for process mining, known as the L* model [10], encompasses five distinct stages to guide the project:

- *Stage 0:* This initial stage involves the planning process and providing a justification for undertaking the project.
- *Stage 1:* Once the project is initiated, the extraction of event data, models, objectives, and pertinent questions is required. This entails obtaining data from various systems, domain experts, and management. It involves understanding the available data for analysis and identifying key questions relevant to the domain.
- *Stage 2:* In this stage, the construction of the control-flow model takes place, which is then linked to the event log. Automated process discovery techniques can be employed for this purpose.
- *Stage 3:* At this stage, the process becomes relatively structured, and the control-flow model can be expanded to include the process mining perspectives such as data, time, and resources.
- *Stage 4:* The models developed in Stage 3 serve as a foundation for operational support. Historical event data is combined with real-time information about ongoing cases. This integrated knowledge can be utilized for interventions, predictions and recommendations.

In summary, the L* life-cycle model comprises the aforementioned stages, encompassing planning and justification, data extraction, control-flow model construction, integrated process modeling, and operational support. It facilitates the extraction of knowledge from historical event data and its application in operational decision-making, prediction, and intervention.

III. Deep Learning and Attention

Deep Learning (DL) [11] is a machine learning field that uses multiple interconnected layers of computation nodes to perform non-linear transformations on data. By adjusting its node parameters, the computational graph can learn patterns that allow it to make classification or regression predictions. This section describes the main DL architectures used for BPM application, focusing on attention mechanisms.

A. Recurrent Neural Networks

RNNs (Recurrent Neural Networks) are a family of neural nets widely used in sequence modeling. These models process an input sequence $x_1, \ldots, x_T$ one element at a time, updating a hidden state h_t that keeps a trace of all the sequence information up to time step t. The hidden state h_T obtained after processing the last input token (x_T) can be used to make predictions about the sequence. For instance, by feeding h_T into a simple fully-connected neural net, one can obtain a sequence classifier. Unfortunately, as many authors have noticed (see, e.g., [12]), RNNs are highly ineffective in learning long-term dependencies between time steps. Long Short-Term Memory (LSTM) alleviate this problem by using gates (usually for input, output and forget data) so that they are able to regulate information flow to and from the model's internal memory. The model can then retain data for more time by learning how and when to open and close these gates properly. Variants of this approach include the GRU (Gated Recurrent Units) architecture, that simplifies the LSTM model by using only two gates, and bi-directional models that process the input sequence using two neural nets: one from right to left and the other in the backward direction [13].

B. Encoder-Decoder Attention

Gated models significantly improve simple RNNs for learning temporal dependencies, but detecting and retaining associations for long time periods is still challenging with these models. This difficulty becomes clear in seq2seq applications such as text translation, which typically require training an architecture with two different neural nets. First, an encoder network processes the input sequence, producing a hidden state h_T. Then, a decoder network conditioned on h_T predicts the output sequence. Unfortunately, it is difficult for the encoder to preserve all the helpful information about the input into a single state h_T. Similarly, it is hard for the decoder to make predictions using a global context vector $c_T = h_T$.

Attention [14] was first introduced to correct the limitation of considering only the last encoder's state, allowing the decoder to access the entire trace of hidden states $h_1, h_2, \ldots, h_T$. At the same time, attention sought to allow the decoder to focus on the most relevant hidden states at each step s of the decoding process. To this end, most attention mechanisms first compute a score α_{st} determining the relevance of h_t for the s-th decoding task and then produce a time-varying context vector c_s computing weighted superposition of the encoder's hidden states. For example, in Bahdanau's additive model [14], the relevance scores are computed by a small neural net. Later, Luong et al. [15] proposed a multiplicative attention mechanism that obtains relevance scores through a (generalized) dot product between a decoder's and encoder's states.

C. Multi-head Attention and the KVQ Abstraction

The Transformer [16] is a type of model that rely on attention mechanisms and have outperformed other deep learning and machine learning methods, especially in sequence prediction tasks [17]. Two key features of Transformers are *multi-head attention* and *positional encoding*. Instead of using a single attention mechanism to compute context vectors for each position, the Transformer uses multiple attention heads that can compute different attention weights and thus learn different relationships among the input data. In addition, positional encoding makes the representation of each input token sensitive to its position by learning and embedding a layer that assigns similar representations to tokens close to each other in the input sequence. Besides stacking attention modules in parallel, the Transformer also stacks attention modules in depth. Indeed, in a standard Transformer, four attention blocks are stacked on each other, and the output of each attention mechanism flows through a position-wise feed-forward net before being used by the next level.

D. Other Attention Mechanisms and Taxonomy

Even if the background material presented above is enough to understand the primary use of attention in BPM, applications in other areas go well beyond sequence modeling. For instance, attention has found wide applications in ML architectures used to learn from graph and image data. As a result, attention mechanisms now come in many different forms and are used at different locations of the deep computational graph. Therefore, it is helpful to rely on a taxonomy to keep track of the latest developments and forecast future applications in BPM.

We can organize current attention models according to four criteria [18]. First, depending on the form of the input, an attention mechanism can be item-wise and location-wise. Item-wise models assume the input data is a pre-segmented sequence of values to which the model can pay attention. Location-wise models, in contrast, deal with data in which there are no explicit parts to attend to; thus, the attention masks' extent is fuzzier. The latter is, for example, the case of image data.

Second, depending on the form of the output, an attention mechanism can be single-output or multi-output. Single-output models encode all the task-relevant information about the input data using a single feature representation. Multi-output models, in contrast, use multiple subspaces to detect different relationships in the input data and thus produce multiple alternative encodings of that data. The latter is, for example, the case of the Transformer's multi-head attention mechanism.

Third, attention mechanisms may have access to different input representations. Depending on the use of a single input, multiple independent inputs, or multiple views of a single input, attention can be classified as *self-attention*, *distinctive attention*, or *co-attention*, respectively. At the same time, attention weights can be computed on input representations at different levels of abstraction, which leads to *hierarchical attention*. In natural language processing, for example, hierarchical attention may operate at the word level and at the sentence level to identify specific and general patterns in the input text.

Finally, an attention mechanism may use all the input values to compute an output or filter out some of them. The first case, known as *soft* or *global attention*, makes the mechanism a continuous deterministic function of its parameters that admits back-propagation learning directly. The second case, known as *hard attention*, includes discrete and stochastic mechanisms that often require devising cumbersome Monte Carlo estimates of the gradient to support learning. The *local attention* mechanism proposed in [15] manages to be sharper than soft attention but computationally more efficient than hard attention by considering a subset of the input values at each time.

IV. Previous Related Surveys

Several recent surveys have explored the use of Machine Learning in Business process monitoring [19] [20] [21], and Deep Learning for Process Mining [3] [7]. In this section, the related surveys are presented, depict their focus topic, approaches and brief conclusions.

Marquez et al. [19] summarized 39 works in the scope of Machine learning methods applied in predictive mon-

itoring of business process, segmenting the methods by their process-awareness and task problems: regression and classification. They concluded that the performance and the accuracy of the methods will highly depend on the attributes of the data rather than the methodology they used to pre-process and process the input data.

Verenich et al. [20] performed a systematic literature review and describe the taxonomy of 25 remaining time prediction studies in the context of business processes and further, they perform an empirical cross-benchmark comparison of 16 different methods across 17 real-life datasets. In this scenario, it was found that LSTM-based methods exhibited superior predictive performance compared to other methods but it is computationally highly demanding. On the other hand, while generative and hybrid techniques provided interpretable models, their predictive accuracy was observed to be significantly lower. This limitation precludes their usage by process analysts.

Teinemaa et al. [21] conducted a review that focused on the standardization of the evaluation methods for the outcome prediction task to ensure valid comparisons between methods. The review examined the taxonomy over 14 methods and included an experimental evaluation of 24 outcome prediction tasks. In the benchmark, it was observed that utilizing the frequency approach to serve the activity attribute as input yields better results compared to sorting the activities by timestamp, this is because the representation of traces are regardless to their lengths.

Rama et al. [7] conducted a comprehensive review on deep learning methods in Process Mining and performed an exhaustive experimental evaluation of ten different methods across twelve public datasets under five different perspectives: input data, predictions, neural network type, sequence encoding, and event encoding. They found that, with few exceptions, using all attributes has a negative impact on the prediction accuracy. Additionally, in the suffix prediction task, the selection of a suitable sampling strategy was found to be dependent on the length of the event log traces, whereas have long of short traces.

Neu et al. [3] presented a broad overview of Deep Learning methods applied in Process Mining. They describe the advantages and disadvantages of the procedural decisions in these approaches from the input encoding to the prediction target. In this paper, the attention mechanisms are mentioned as part of the network architecture section, but their characteristics and applications were not delved into. Further, multiple research lines are proposed: *Prediction to action*, *Leveraging domain knowledge process mining*, *input variables* and *causality and interpretability*. In summary, the authors suggest that the methods should include human domain knowledge and provide transparency to the decision maker about the inner behavior of the model to have information of how the predictions were made.

There are also several reviews on the use of attention mechanisms in other fields, such as: image processing [4], audio recognition [5], image captioning [6], etc. But none of these studies have explored in depth the use of attention mechanisms for process mining.

V. Methodology - Systematic Literature Review

To carry out this study, a systematic literature review was conducted, following Wohlin's recommendations [22]. Furthermore, we conducted literature filtering using the PRISMA flow methodology [23].

1) Define research questions.
2) Define search keywords for relevant articles in academic databases.
3) Set inclusion and exclusion criteria to filter the found articles and leave only those relevant to answer the research questions.
4) Finally, an additional manual search without the keywords, to expand the set of articles as some publication may have a keyword synonym.

A. Research questions

The objective of this review is to identify both the *attention mechanisms* and their *applications in process mining* to date. Thus, the research questions are:

- (RQ1) For which applications are AM used?
- (RQ2) Which kinds of AM are used?
- (RQ3) How are AM combined with other techniques?

B. Search strings

The search as conducted on the main academic databases for computer science: Web of Science[1], IEEE Explore[2], ACM Digital Library[3], ScienceDirect[4] and SpringerLink[5].

The search strings were "business process*" (to restrict scope to Process Mining) and "event log*" (to expand the search); and "attention*" (to specify studies on attention mechanisms). Database queries were done on title, keywords, and abstract; full text queries were left out because a trial run yield many false positives. Duplicates (works indexed in more than one database) were counted as one, and for works with more than one version, only the peer-review final publication was included. Table I shows the search strings and databases, as well as results filtered by the title over the total. Furthermore, the PRISMA flow diagram is depicted in Figure 1, outlining the four primary filtering results by each step: title-based selection, screening, eligibility based on inclusion/exclusion criteria, and the synthesis of studies identified through alternative methods.

C. Inclusion and exclusion criteria

Finally, inclusion and exclusion criteria were applied to determine the final set of articles. These criteria were:

- Inclusion Criteria:

[1] https://www.webofknowledge.com/
[2] https://ieeexplore.ieee.org/
[3] https://dl.acm.org/
[4] https://www.sciencedirect.com/
[5] https://link.springer.com/

TABLE I
SEARCH STRINGS AND DATABASES (ARTICLES THAT PASSED THE TITLE FILTERING VS TOTAL NUMBER OF ARTICLES RETURNED BY THE QUERY

Database	Search string	Results
Web of Science	Title: Business Process* AND Title: Attention mechanism*	2/2
Web of Science	Abstract: Business Process* AND Abstract: Attention mechanism*	14/151
Web of Science	Title: Event log* AND Title: Attention mechanism*	0/0
Web of Science	Abstract: Event log* AND Abstract: Attention mechanism*	8/71
Web of Science	Topic: Business Process* AND Topic: Attention mechanism* AND Topic: Neural network*	5/22
Science Direct	Find articles with these terms: "Business process" AND "attention mechanism"	17/83
Science Direct	Find articles with these terms: "Event log" AND "Attention mechanism"	16/46
IEEE Explore	All metadata: "Business process" AND All Metadata: "attention mechanism"	5/10
IEEE Explore	All Metadata: "event log*" AND All Metadata:"attention mechanism"	6/7
ACM	Applied Filters: Research articles, [All: "business process*"] AND [All: "attention mechanism*"]	2/13
ACM	Applied Filters: Research articles, [All: "event log*"] AND [All: "attention mechanism"]	4/9
Scopus	Title, Abstract & keywords: "business process*" AND "attention mechanism*"	17/22
Scopus	Title, Abstract & keywords: "event log*" AND "attention mechanism*"	8/19

IN1: The study focuses on business process and presents a method that has been deployed and assessed.

IN2: The study uses an attention mechanism.

- Exclusion Criteria:

EX1: The study is not published in English (which may pose challenges in understanding key concepts).

EX2: The study is electronically unavailable, or not not freely accessible through standard university libraries services.

EX3: The approach is not related to process mining (which is the main focus of this review).

The search strings on the databases yield 73 studies[6].

D. Manual search

Manual search with the keyword synonyms "system logs*," "process mining" and "predictive process monitoring" yield fifteen more studies, of which only one was added to the review set.

The results of both database and manual searches were filtered using the inclusion and exclusion criteria, resulting in sixteen studies passing this stage. Most rejections were due to duplication, lack of direct relevance or explicit mention of Process Mining, or a lack of an attention mechanism as a component in their architecture. Table II presents the sixteen selected studies that were ultimately reviewed.

VI. RESULTS & DISCUSSION

This section reviews the studies to answer the three research questions: Subsection VI-A for RQ1 (classifies them according to their application, tasks they seek to solve, general model architecture, and each application and task); Subsection VI-B for RQ2 (the types of attention mechanisms found in the surveyed studies); and Subsection VI-C for RQ3 (their part within the neural network model).

[6]Search results are located in a public Google Drive repository https://drive.google.com/drive/folders/1LUH037mmC6JtLReQ_G8KpUohiHtbeUsW?usp=sharing

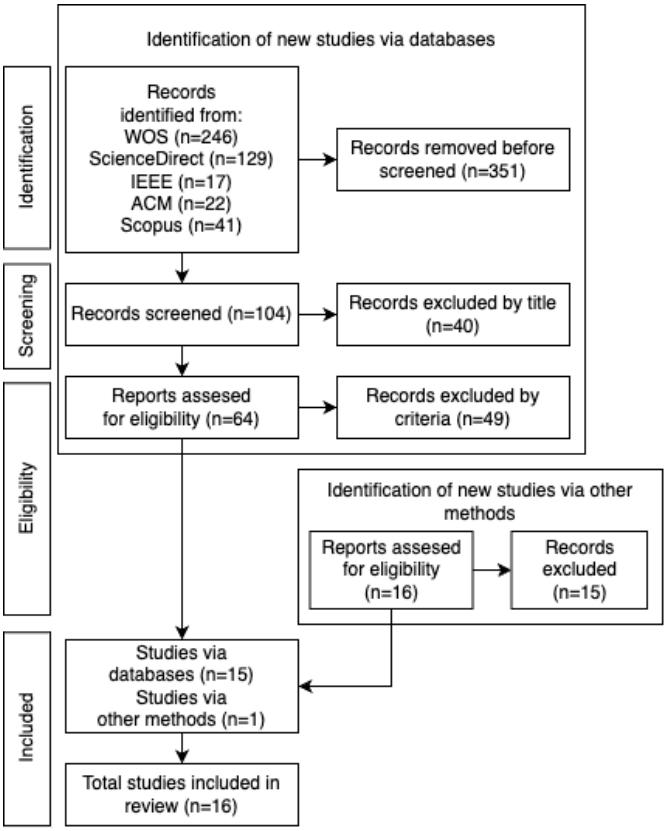

Fig. 1. PRISMA flow diagram: Research studies, identified through databases and other sources, underwent filtering based on screening and inclusion/exclusion criteria.

A. RQ1: Applications of attention mechanisms for PM

In Table II provides a summary of all the works and their corresponding applications. Three research lines can be identified: (1) Process prediction, (2) Data representation, and (3) Anomaly detection, in decreasing number of papers.

1) Process prediction: Process prediction is a well-studied task where deep learning methods have managed to far outperform traditional Machine Learning models and statistical methods [3] [7]. The tasks that are sought to be

TABLE II
OVERVIEW REGARDING THE APPLICATIONS OF THE REVISED METHODS THAT INCORPORATE ATTENTION MECHANISMS

Study	Architecture	Application			
		Process Prediction		Data Representation	Anomaly Detection
		Classification/Regression	Interpretability		
Krahsic & Franczyk (2021) [24]	LSTM Autoencoder	-	-	-	✓
Wickramanayake et al. (2021) [25]	LSTM	Next activity	✓	-	-
Jalayer et al. (2020) [26]	Bi-LSTM E/D	Next activity	-	-	-
Jalayer et al. (2022) [27]	Bi-LSTM HAM E/D	Next activity	✓	-	-
Wang et al. (2019) [28]	Bi-LSTM E/D	Outcome	-	-	-
Heinrich et al. (2021) [29]	KVP Attention	Next activity	-	-	-
Weytjens & Weerdt (2020) [30]	LSTM	Outcome	-	-	-
Moon et al. (2021) [31]	POP-ON Transformer	Next activity	-	-	-
Phillip et al. (2020) [32]	Transformer	Next activity	-	-	-
Bukhsh et al. (2021) [33]	Transformer	Next activity/Timestamp	-	-	-
Rivera-Lazo & Ñanculef (2022) [34]	Transformer	Next Multi-attribute	-	-	-
Guzzo et al. (2021) [35]	LSTM Autoencoder	-	-	✓	-
Stierle et al. (2021) [36]	Gated GNN	-	-	✓	-
Harl et al. (2021) [37]	Gated GNN	-	-	✓	-
Hnin et al. (2019) [38]	LSTM	Next attribute + Outcome	-	-	-
Lin et al. (2019) [39]	LSTM	Next attribute	✓	-	-

solved are specifically classification and regression, that is, it is sought to be able to predict the next event, when it will be triggered or what the output of that trace will be. To measure these models' performance in both tasks, the set metrics are *accuracy* for classification tasks, *AUC_ROC* for outcome prediction, and *MAE* for regression. Also, it is common to use BPM competency datasets among other event logs such as Helpdesk[7], BPIC2012[8], BPIC2013[9], and BPIC2017[10].

Most studies use recurrent networks of type LSTM [25] [30] [38] or Bi-LSTM [26] [28] [27]. Recent publications have opted for parallelizable models for computational efficiency issues such as Transformer encoders [31] [32] [33] [34]. On the other hand, Heinrich et al. [29] uses a key-value-predict attention network.

In addition to the increase in performance and accuracy provided by using attention mechanisms in neural networks, some works seek to grant interpretability to the model, namely Wickramanayake et al. [25] and Jalayer et al. (2022) [27], which aims to relate output with the weights of the attention layer according to the input, as shown in Fig. 2.

Regarding the task of predicting the next activity, Table III shows that the studies achieving the highest accuracy are those that incorporate not only the activity, but also categorical and continuous attributes as input, as they provide relevant information to the model. The study conducted by Hnin et al. in 2019 [38] did not report any results.

There are only two studies that consider **predicting the time when the next activity is going to be triggered**: Bukhsh et al. [33], and Rivera-Lazo & Ñanculef [34], which proposed a multi-task approach that attempt to predict the next activity, resource and timestamp within the same model.

[7]https://data.4tu.nl/repository/uuid:0c60edf1-6f83-4e75-9367-4c63b3e9d5bb
[8]https://doi.org/10.4121/uuid:3926db30-f712-4394-aebc-75976070e91f
[9]https://www.win.tue.nl/bpi/doku.php?id=2012:challenge
[10]https://data.4tu.nl/articles/BPI_Challenge_2017/12696884

Lin et al. [39] proposed a LSTM encoder-modulator-decoder architecture, where the modulator layer allows to extract model insights using an implicit attention mechanism; this layer infers an alignment weight vector of each activity and attribute produced by an element-wise multiplication of all encoder representations. Lin et al. also present results within suffix generation task by using the same architecture and iteratively feeding the model with the predicted tokens.

Finally, in the **outcome classification task**, Table IV shows Wang et al. [28] and Weytjens & Weerdt [30] results; these methods can not be compared as they have different evaluation settings, such as the length of the prefix which directly affects the AUC metric precision.

2) Data Representation: Data representation is a important task for analysts to gain a better understanding of how complex processes operate. Therefore, the use of attention mechanisms that assign importance values to process elements or to the data representations is a common approach in this area of research.

Three studies were identified, two of them proposed the use of a neural network called *Gated Graph Neural Network* to obtain a graph representation where each node corresponds to an activity and its value represents the level of relevance within the context. These studies are Harl et al. (2020) [37] and Stierle et al. (2021) [36]. Harl et al. [37] additionally included the outcome classification task, and reported the process mapped by the scores of each activity with respect to the outcome as shown in Figure 3.

Guzzo et al. [35] generate a representation of the traces as an embedded vector through an unsupervised method and an attention layer using data from the different perspectives of the process, refers as control-flow perspective, data perspective and organizational perspective. To validate their method, they test these representations on two prediction tasks: trace classification and clustering, outperforming the current state-of-the-art methods.

#Event	Activity	Resource	Action	EventOrigin	Lifecycle:transition
1	A_Create_Application	User_1	Created	Application	complete
2	A_Submitted	User_1	statechange	Application	complete
3	W_Handle_leads	User_1	Created	Workflow	schedule
4	W_Handle_leads	User_1	Deleted	Workflow	withdraw
5	W_Complete_application	User_1	Created	Workflow	schedule
6	A_Concept	User_1	statechange	Application	complete
7	W_Complete_application	User_17	Obtained	Workflow	start
8	W_Complete_application	User_17	Released	Workflow	suspend
9	W_Complete_application	User_38	Obtained	Workflow	resume
10	A_Accepted	User_38	statechange	Application	complete
11	O_Create_Offer	User_38	Created	Offer	complete

Predicted next activity: o_created

Fig. 2. This is a predicted partial trace from BPIC 2017 evaluated in Jalayer et al. [27]. There are 11 events and each event has 5 attributes, i.e., Activity, Resource, Action, Event Origin and, Lifecycle: transition. The blue color indicates the amount of attention to an event in a trace and the red color indicates the amount of attention to each attribute to create the feature vector of that event. The darker the color, the more attention has been paid.

TABLE III
BASELINES COMPARISON FOR THE NEXT ACTIVITY WITH INPUTS AS ACTIVITY (A), RESOURCE (R), TIMESTAMP (T) AND CATEGORICAL ATTRIBUTES (C).

Author	Architecture	Input	Percentage	Prefix	Accuracy			
					Helpdesk	*BPIC 2012*	*BPIC 2013*	*BPIC 2017*
Wickramanayake et al.	LSTM	A	70/30	1	-	-	0.707	-
Jalayer et al. (2020)	bi-LSTM	A	80/20	2	0.833	0.816	-	-
Jalayer et al. (2022)	HAM-net bi-LSTM	C	80/20	1	0.844	0.868	-	0.929
Heinrich et al.	KVP attention	A,C,T	80/20	1	0.857	0.855	0.666	-
Lin et al.	LSTM	A,C	70/20/10	5	0.916	**0.974**	-	**0.974**
Moon et al.	POP-ON	Custom A	80/20	1	0.755	0.794	**0.798**	-
Phillip et al.	Transformer	A	66/33	2	-	-	0.707	-
Bukhsh et al.	ProcessTransformer	A	60/20/20	1	0.856	0.852	0.621	-
Rivera-Lazo & Ñanculef	MA-Transformer	A,R,T	80/20	1	**0.924**	0.892	-	0.823

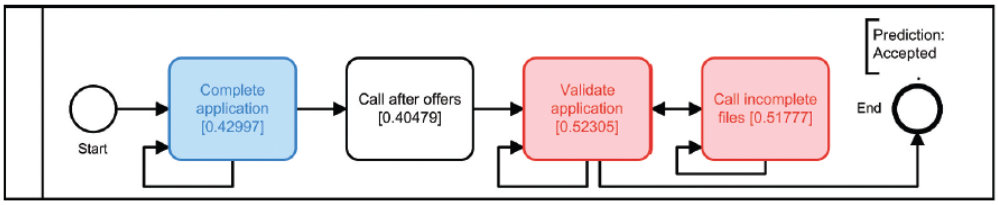

Fig. 3. Visualization of the process model with the activity scores according to the outcome prediction based on the Harl et al. (2020) [37] method. The activity colors are arranged in order of increasing relevance from low to high, starting with black/white, followed by blue and red.

TABLE IV
AUC RESULTS OF THE OUTCOME PREDICTION TASK.

Dataset	Wang et al.		Weytjens & Weerdt	
	Prefix	AUC	Prefix	AUC
BPIC 2012 accepted	40	0.71	35	0.79
BPIC 2012 declined	40	0.64	5	0.75
BPIC 2012 canceled	40	0.73	5	0.75
BPIC 2017 accepted	20	0.84	47	0.93
BPIC 2017 declined	20	0.83	47	0.91
BPIC 2017 canceled	20	0.82	47	0.92
Traffic fines	10	0.73	6	0.77

3) Anomaly detection: Only one paper is related to this application. Krahsic & Franczyk [24] evaluate a semi-supervised deep learning method to classify anomalous traces from three datasets, two of them generated synthetically by the tool called "PLG2"[11] proposed by Burattin et al. [40]; and the third dataset is a product of the 2013 BPIC contest.

B. RQ2: Types of attention mechanism used for PM

Table V categorizes the reviewed articles according to the four-fold taxonomy presented in Section III. Regarding the input representation, most studies use *self-attention* to process a single input sequence. Exceptions to this are Jalayer et al. (2022) [27] and Guzzo et al. [35], which use *hierarchical self-attention* to obtain representations at several abstraction levels. Rivera-Lazo and Ñanculef [34] also evaluate variants of attention mechanisms that take multiple attributes as input, which is known as "co-attention".

[11]https://github.com/delas/plg

Considering the output representation, most applications rely on primitive attention mechanisms, capable of producing a single representation per time step. We note four exceptions to this pattern [34] [31] [33] [32], all of which exploit a multi-head attention approach inherited from using a Transformer encoder to represent the input sequence before making a prediction.

Concerning the sharpness/softness of the attention mechanisms and shape of the input features, all the surveyed applications employ global and item-wise attention, respectively. Given that hard and local attention mechanisms are thought to provide higher interpretability, we deem their exploration in the context of Process Mining applications to be worthwhile despite the technical challenges involved in their implementation.

C. RQ3: Neural network architectures with attention used for PM

The results of the review indicate that all of the methods position the attention mechanism after an encoder layer, which generates a vectorial representation of the input. The reviewed methods can be classified into four distinct architecture types based on their choice of encoder method.

1) *Networks with RNNs:* In recurrent networks such as LSTM, GRU and Bi-LSTM, the attention mechanism can be positioned in both encoder and encoder-decoder architectures to generate a context vector that maps to each input sequence at each recurrent step. The works proposing recurrent encoders include [25] [30] [38] [39]. Moreover, the encoder-decoder architectures have been proposed by [24] [26] [27] [28] [35].
2) *Transformers:* The methods that are being considered involve incorporating an attention layer in both the encoder and decoder. However, related works such as Moon et al. [31], Phillip et al. [32], Bukhsh et al. [33], and Rivera-Lazo & Ñanculef [34] have only evaluated the encoder Transformer architecture.
3) *Graph Neural Networks:* Harl et al. [37] and Stierle et al. [36] used a graph neural network in conjunction with an attention mechanism, which allows them to extract the relevance of the tokens in the sequence related to an output.
4) *Gated CNN:* Heinrich et al. (2021) [29] propose to use a *Gated Convolutional Neural Network* with a Key-value-predict attention layer. This architecture is a non-recurrent network alternative to capture long-term dependencies while avoiding sequential operations for better parallelizability.

In summary, the most used architectures are those that include recurrent networks, and specifically LSTM, which produce as output a continuous representation that serves directly as input to the attention layer. Only a few papers explore other architectures, like *Transformers*, *Graph Neural Networks* or *Gated Convolutional Neural Networks*.

VII. Future Challenges & Conclusion

By considering the arising of attention mechanisms in Process Mining, this study has outlined published literature to describe their application, design and findings as a result of a systematic literature review of these mechanisms in the scope of Process Mining. Filtered by a set of inclusion/exclusion criterias, the initial search of 73 papers were reduce to sixteen studies, which we describe and classified by their applications, mechanism types and position in the area. Further, proposals were compared by means of metrics and designs, obtaining relevant information to continue the evolution of these research lines, such as: Key findings are: (1) The main applications of attention-based methods in process mining are process prediction, data representation, and anomaly detection. Most of the research has focused on the process prediction task, and those that included attention mechanisms, such as Transformer and LSTM encoder-decoder architectures, have demonstrated greater accuracy compared to other approaches, when predicting the next event or process outcome. (2) Additionally, most of these methods provide insights and transparency to decision makers, as the internal attention of the method's input in relation to the output prediction can be visualized. (3) The majority of taxonomy criteria for the methods are similar, with only three papers differing in their treatment of inputs by using hierarchical self-attention to segment between activity and other event attributes and co-attention to fuse multiple attributes. Additionally, the Transformers architecture naturally considers to process multiple outputs. (4) The majority of the methods incorporate the attention mechanism following the encoder to produce an attention vector representation that selectively focuses on specific input tokens. (5) Gradually aggregating attributes that characterize the events into the model provides valuable information to enrich the model's input and gain insight into how they influence the target prediction. However, due to different experimental settings, it can be difficult to compare metrics between methods.

Further, multiple research opportunities have been identified: (1) *Neural Network Architecture:* To improve not only the prediction accuracy but also the identification of patterns in single or multiple attributes fusion, it is recommended to explore other attention-based taxonomy criteria, such as co-attention, local attention, and multi-output. Additionally, further experimentation with other types of neural network architectures such as Graph Neural Networks, Gated-CNNs, etc. that create representations from different input data structures should also be considered. (2) *Applications:* Generating synthetic data is a well-known application in Process Mining and other areas of machine learning. It is used to create new examples when there is insufficient data or to complete a sequence, image, audio, or other type of data. For example: by utilizing generative models that include attention mechanisms with a multi-modal approach, it may be possible to combine business process representations,

TABLE V
SUMMARY OF THE TAXONOMIES OF THE ATTENTION MECHANISMS. THE REVISED PAPERS ARE CLASSIFIED BY THE FOUR CRITERIA PRESENTED BY NIU ET AL. [18].

Study	Softness	Forms	Input	Output
Krahsic & Franczyk	Soft/global	Item-wise	Self-attention	Single
Wickramanayake et al.	Soft/global	Item-wise	Self-attention	Single
Jalayer et al. (2020)	Soft/global	Item-wise	Self-attention	Single
Jalayer et al. (2022)	Soft/global	Item-wise	Hierarchical	Single
Wang et al.	Soft/global	Item-wise	Self-attention	Single
Heinrich et al.	Soft/global	Item-wise	Self-attention	Single
Weytjens & Weerdt	Soft/global	Item-wise	Self-attention	Single
Moon et al.	Soft/global	Item-wise	Self-attention	Multi
Phillip et al.	Soft/global	Item-wise	Self-attention	Multi
Bukhsh et al.	Soft/global	Item-wise	Self-attention	Multi
Rivera-Lazo & Ñanculef	Soft/global	Item-wise	Self and Co-attention	Multi
Guzzo et al.	Soft/global	Item-wise	Hierarchical	Single
Stierle et al.	Soft/global	Item-wise	Self-attention	Single
Harl et al.	Soft/global	Item-wise	Self-attention	Single
Hnin et al.	Soft/global	Item-wise	Self-attention	Single

organizational graph representations, and temporal patterns to create new scenarios from the three Process Mining perspectives. For example, such models could be used to complete a sequence subject to certain temporal restrictions or to create new work teams. (3) *Interpretability:* Lastly, it remains a challenge to further explore how the insights from the internal data of the model can be utilized as a structure source of valuable information to be interpreted by decision makers and process analysts, potentially leading to a better understanding of the process and the business.

ACKNOWLEDGEMENTS

This work has been partially supported by ANID under grant PIA/APOYO AFB220004 (CCTVal).

REFERENCES

[1] S. Kraus, S. Durst, J. J. Ferreira, P. Veiga, N. Kailer, and A. Weinmann, "Digital transformation in business and management research: An overview of the current status quo," *International Journal of Information Management*, vol. 63, p. 102466, 2022.

[2] W. Van Der Aalst, "Process mining," *Association for Computing Machinery*, vol. 55, no. 8, 2012.

[3] D. A. Neu, J. Lahann, and P. Fettke, "A systematic literature review on state-of-the-art deep learning methods for process prediction," *Artificial Intelligence Review*, vol. 55, no. 2, pp. 801–827, Mar. 2021.

[4] S. Ghaffarian, J. Valente, M. van der Voort, and B. Tekinerdogan, "Effect of attention mechanism in deep learning-based remote sensing image processing: A systematic literature review," *Remote Sensing*, vol. 13, no. 15, p. 2965, Jul. 2021.

[5] E. Lieskovská, M. Jakubec, R. Jarina, and M. Chmulík, "A review on speech emotion recognition using deep learning and attention mechanism," *Electronics*, vol. 10, no. 10, p. 1163, May 2021.

[6] Z. Zohourianshahzadi and J. K. Kalita, "Neural attention for image captioning: review of outstanding methods," *Artificial Intelligence Review*, vol. 55, no. 5, pp. 3833–3862, Nov. 2021.

[7] E. Rama-Maneiro, J. C. Vidal, and M. Lama, "Deep learning for predictive business process monitoring: Review and benchmark," *CoRR*, 2020.

[8] M. Dumas, M. La Rosa, J. Mendling, H. A. Reijers *et al.*, *Fundamentals of business process management*. Springer, 2013, vol. 1.

[9] W. M. P. van der Aalst, *Process Mining: Data Science in Action*. Springer, 2016.

[10] W. e. a. van der Aalst, "Process mining manifesto," in *Business Process Management Workshops*. Springer, 2012, pp. 169–194.

[11] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[12] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

[13] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *arXiv*, 2014.

[14] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[15] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *EMNLP*. ACL, Sep. 2015, pp. 1412–1421.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017.

[17] A. Gillioz, J. Casas, E. Mugellini, and O. A. Khaled, "Overview of the transformer-based models for nlp tasks," in *FedCSIS*, 2020, pp. 179–183.

[18] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, Sep. 2021.

[19] A. E. Márquez-Chamorro, M. Resinas, and A. Ruiz-Cortés, "Predictive monitoring of business processes: A survey," *IEEE Transactions on Services Computing*, vol. 11, no. 6, pp. 962–977, 2018.

[20] I. Verenich, M. Dumas, M. L. Rosa, F. M. Maggi, and I. Teinemaa, "Survey and cross-benchmark comparison of remaining time prediction methods in business process monitoring," *CoRR*, vol. abs/1805.02896, 2018.

[21] I. Teinemaa, M. Dumas, M. L. Rosa, and F. M. Maggi, "Outcome-oriented predictive process monitoring: Review and benchmark," *CoRR*, vol. abs/1707.06766, 2017.

[22] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén, *Systematic Literature Reviews*. Springer, 2012, pp. 45–54.

[23] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hróbjartsson, M. M. Lalu, T. Li, E. W. Loder, E. Mayo-Wilson, S. McDonald, L. A. McGuinness, L. A. Stewart, J. Thomas, A. C. Tricco, V. A. Welch, P. Whiting, and D. Moher, "The prisma 2020 statement: An updated guideline for reporting systematic reviews," *International Journal of Surgery*, vol. 88, p. 105906, 2021.

[24] P. Krajsic and B. Franczyk, "Semi-supervised anomaly detection in business process event data using self-attention based classification," *Procedia Computer Science, KES*, vol. 192, pp. 39–48, 2021.

[25] B. Wickramanayake, Z. He, C. Ouyang, C. Moreira, Y. Xu, and R. Sindhgatta, "Building interpretable models for business process prediction using shared and specialised attention mechanisms," *CoRR*, vol. abs/2109.01419, 2021.

[26] A. Jalayer, M. Kahani, A. Beheshti, A. Pourmasoumi, and H. R. Motahari-Nezhad, "Attention mechanism in predictive business process monitoring," in *EDOC*. IEEE, Oct. 2020.

[27] A. Jalayer, M. Kahani, A. Pourmasoumi, and A. Beheshti, "HAM-net: Predictive business process monitoring with a hierarchical attention mechanism," *KBS*, vol. 236, p. 107722, Jan. 2022.
[28] J. Wang, D. Yu, C. Liu, and X. Sun, "Outcome-oriented predictive process monitoring with attention-based bidirectional lstm neural networks," in *2019 IEEE International Conference on Web Services (ICWS)*, 2019, pp. 360–367.
[29] K. Heinrich, P. Zschech, C. Janiesch, and M. Bonin, "Process data properties matter: Introducing gated convolutional neural networks (GCNN) and key-value-predict attention networks (KVP) for next event prediction with deep learning," *Decision Support Systems*, vol. 143, p. 113494, 2021.
[30] H. Weytjens and J. D. Weerdt, "Process outcome prediction: CNN vs. LSTM (with attention)," in *Business Process Management Workshops*. Springer International Publishing, 2020, pp. 321–333.
[31] J. Moon, G. Park, and J. Jeong, "POP-ON: Prediction of process using one-way language model based on NLP approach," *Applied Sciences*, vol. 11, no. 2, p. 864, Jan. 2021.
[32] P. Philipp, R. Jacob, S. Robert, and J. Beyerer, "Predictive analysis of business processes using neural networks with attention mechanism," in *ICAIIC*. IEEE, Feb. 2020.
[33] Z. A. Bukhsh, A. Saeed, and R. M. Dijkman, "Processtransformer: Predictive business process monitoring with transformer network," *CoRR*, vol. abs/2104.00721, 2021.
[34] G. Rivera-Lazo and R. Ñanculef, "Multi-attribute transformers for sequence prediction in business process management," in *Discovery Science*. Springer, 2022, pp. 184–194.
[35] A. Guzzo, M. Joaristi, A. Rullo, and E. Serra, "A multi-perspective approach for the analysis of complex business processes behavior," *Expert Systems with Applications*, vol. 177, p. 114934, 2021.
[36] M. Stierle, S. Weinzierl, M. Harl, and M. Matzner, "A technique for determining relevance scores of process activities using graph-based neural networks," *Decision Support Systems*, vol. 144, p. 113511, May 2021.
[37] M. Harl, S. Weinzierl, M. Stierle, and M. Matzner, "Explainable predictive business process monitoring using gated graph neural networks," *Journal of Decision Systems*, vol. 29, no. sup1, pp. 312–327, Jun. 2020.
[38] T. Hnin and K. K. Oo, "Attention based lstm with multi tasks learning for predictive process monitoring," *WCSE*, p. 165 – 170, 2019.
[39] L. Lin, L. Wen, and J. Wang, "MM-pred: A deep predictive model for multi-attribute event sequence," in *Proceedings of the 2019 SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics, May 2019, pp. 118–126.
[40] A. Burattin, "PLG2: multiperspective processes randomization and simulation for online and offline settings," *CoRR*, vol. abs/1506.08415, 2015.