PH.D. CONFIRMATION OF CANDIDATURE

Explaining Deep Learning-based Process Predictions

PhD Confirmation Report

School of Information Systems

Bemali Wickramanayake
(Student ID: n10915109)

Submitted in partial fulfillment of the requirement for the degree of
IF49 Doctor of Philosophy

Supervisors:
Dr. Chun Ouyang
Dr. Caterina Moreira
A. Prof. Yue Xu

Thursday 11th August, 2022

Contents

Research Project

0.1 Proposed Thesis Title and Type

Proposed Thesis Title: Explaining Deep Learning-based Process Predictions
Proposed Thesis Type: Traditional Thesis by Monograph

0.2 Proposed Supervisory Team and their Credentials

Principal Supervisor: Dr. Chun Ouyang
Dr. Ouyang is a senior lecturer in the School of Information Systems and has supervised four PhD students to completion. She is an active and well-established researcher and the world top 21^{st} most cited scholar in Process Mining (according to Google Scholar). Built upon her expertise in process-oriented data mining, she has developed strong research interest in explainable predictive process analytics and submitted an ARC Discovery Project on the topic as the lead investigator. She is currently supervising five PhD students as the principal supervisor and one PhD student as an associate supervisor.

Associate Supervisor: Dr. Catarina Pinto Moreira
Dr. Moreira is a senior lecturer in the School of Information Systems and holds the role of Deputy HDR Academic Lead. She is a Computer Scientist and passionate in investigating machine learning / deep learning models, and innovative human interactive probabilistic models for explainable AI. She has submitted a grant proposal on Persuasive and Causal Probabilistic Models for Explainable AI for ARC Discovery Early Career Researcher Award (DECRA). She is currently supervising two PhD students as the principal supervisor and four PhD students as an associate supervisor.

Associate Supervisor: A/Prof. Yue Xu
A/Prof. Xu is from the School of Computer Science and has many years of research experience in data mining and text mining, including leading two ARC grants as a CI and three CRC grants as a senior researcher. She also has considerable research expertise in the areas of pattern and association mining and natural language generation. She has supervised twelve PhD students to completion and is the principal supervisor of seven current PhD students.

0.3 Acknowledgement for existing Funding and Scholarships

This PhD is funded by the following scholarships offered by QUT.

- Science and Engineering Faculty Scholarship (2021-2024)
- Centre for Data Science Top Up Scholarship (2021-2024)

Chapter 1

Introduction

A business process is a collection of interdependent events, activities and decisions, often involving different actors and objects, which will lead to an outcome that meets an organizational goal [59]. Process mining is a field of studies that aims at construction of models that explain the behaviour of a process based on process execution event logs (or event logs for short) [63]. Predictive process analytics (PPA) is an emerging sub domain in the area of process mining. It aims to build the capability for predicting future status of an ongoing process instance, that will help address business purposes such as improving lead time and reducing overall process cost via interacting with and controlling the process execution at runtime and optimizing the process design. In PPA the mostly studied prediction targets are to predict *next event*, *remaining process execution time*, and *process outcome*.

Process predictions are mainly derived from either with process model-based approaches (called 'white-box' approaches in PPA) [4, 13] or machine learning-based approaches [68]. In this research we are interested in the machine learning-based approaches, due to their predictive performance, capability of supporting online predictive process monitoring, and less dependency upon the domain knowledge. Amongst the machine learning-based approaches in PPA, deep learning-based models are gaining more and more attention due to their ability of handling high dimensional feature spaces such as event log data. Also, deep learning-based models are proven to be superior in predictive performance as well [28, 68, 70].

However, deep learning-based models are considered as 'black-boxes' given their sophisticated internal representation and computational complexity. The opaqueness of a predictive model means that the humans cannot understand how the model makes the decisions, and which criteria it considers as important in making those decisions. This lack of transparency results in ambiguity about the validity and fairness of the decisions that are made by the underlying model. Explainable AI (XAI) attempts to fill this gap between the human understanding and automated decision making [20]. The need for explainable AI is motivated by the increased automation of decision making resulting in ambiguity about the decisions for the users who consume them. Further, the regulations such as GDPR stand that users have the fundamental right of knowing how a decision that affects them was arrived at [12].

In XAI research, a black-box model can be explained using a simpler surrogate model to approximate the black-box *(post-hoc techniques)* [20] or using internal properties of the black-box itself *(intrinsic techniques)*. Although the use of surrogate/post-hoc techniques are quite popular in the domain of XAI, one common problem all those techniques have is the explanation fidelity. That is, the explanations need to be tested to confirm if they are truthful to the prediction model. Whereas the intrinsic techniques aim to build interpretability within the internal structure of a prediction model, hence the explanation fidelity is not an issue. However, a key challenge faced by intrinsic techniques is the incorporation of interpretation mechanisms into the sophisticated internal structure of a prediction model without introducing much disruption to the model's performance. It is also worth noting that due to the complicated nature of deep-learning models, by developing intrinsic techniques, we are interested in extracting relevant and useful explanations, rather than making such a model fully transparent.

In the domain of predictive process analytics, the model explainability has a potential to improve the efficiency and effectiveness of the decisions that are made by the predictive systems [20], in particular for

process inspection and optimization. This research is motivated by the fact that the more deep learning techniques have been used to build process predicative capabilities, the more important it is to be able to explain those deep learning-based process prediction models to humans, e.g., for them to inspect the predictive model and the underlying process. Further, the research focuses on the intrinsic techniques of explaining deep learning-based process prediction models because of the faithfulness of those techniques to the prediction model.

Hence, the proposed project will address the topic of intrinsic explainable techniques for deep learning-based process prediction models, with the objective of developing explainability techniques that would suit a certain purpose for applying the extracted explanations. The literary background for the proposed project is laid out in Chapter 2 - Literature Review. Based on the literature review, the research questions are presented in Chapter 3, and research objectives and research plan to address the research questions are proposed in Chapter 4. Chapter 5 describes the progress to date on the project since the preliminary proposal for this project (i.e., Stage 2 proposal) was approved. Chapter 6 discusses the future work to be carried out and expected research timeline till completion.

Chapter 2

Literature Review

In this section the existing literature related to Predictive Process Analytics (PPA) and Explainable AI (XAI) is studied to lay a solid theoretical background that would help identifying and addressing potential research gaps in the area of explainable predictive process monitoring. The review of literature studies existing PPA approaches underpinned by machine learning (including deep learning) techniques and to tackle different process prediction problems, including the state-of-the-art based on benchmark studies. Further, investigating model explainability techniques and their applications in relevance, and identifying evaluation metrics for model explainability are also key objectives in this review.

2.1 Predictive Process Analytics

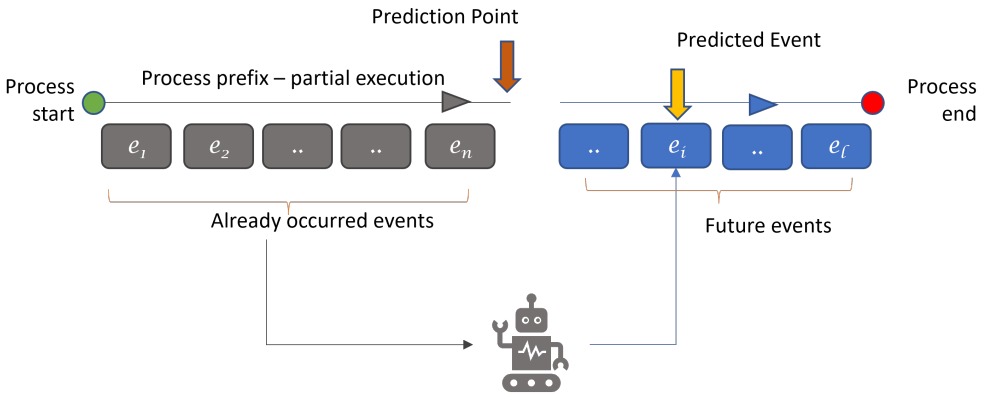


Figure 2.1: Representation of Predictive Process Analytics

Use of Machine Learning-based predictive analytics techniques are gaining popularity in the domain of process mining, due to the enhanced predictive performance and requiring minimal process and domain knowledge as opposed to exploratory process analytics based techniques [4, 13]. The fundamental prediction problem in predictive process analytics is focused on predicting the next event [40] of an ongoing process trace, that eventually leads to the prediction of remaining time to complete the process [68] and the process outcome [59].

Prediction of next event can be a combination of classification and regression problems. Classification is used to predict the activity label (and/ or associated categorical event attributes) and regression is often used to predict the timestamp of the next event. Prediction of remaining time of an ongoing process trace is a regression problem and predicting process outcome can be framed as a classification problem.

Traditional machine learning techniques used for next activity/ event prediction span across a wide

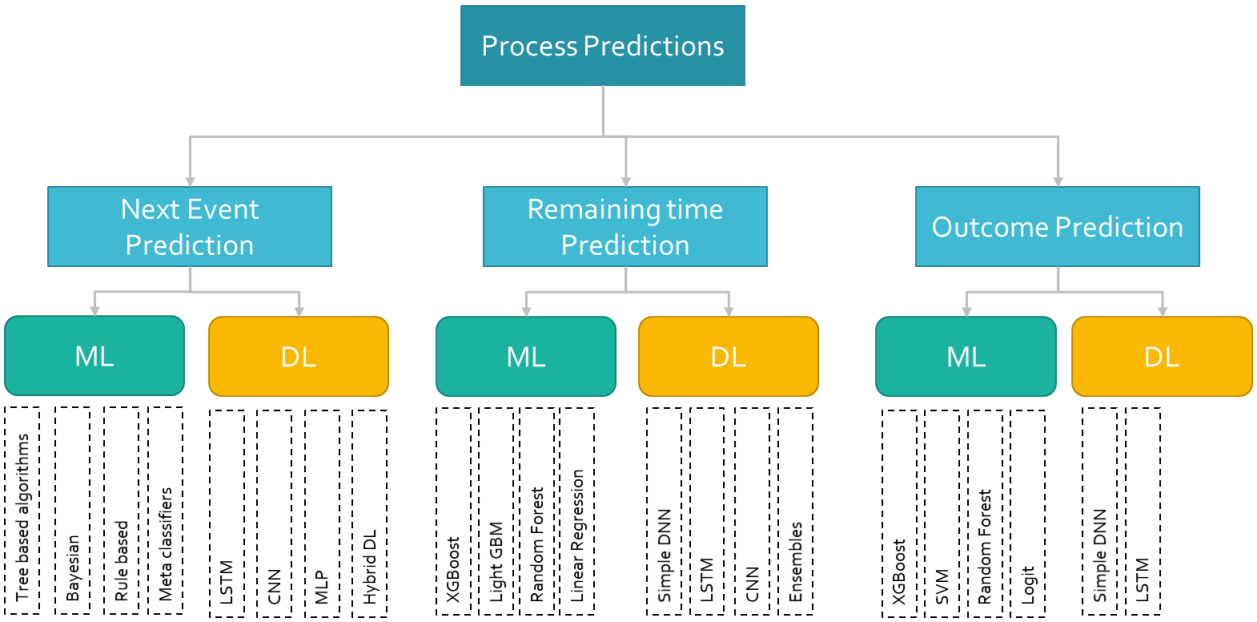


Figure 2.2: Most common Process Prediction Problems and Predictive Techniques used

range of classification algorithms which include Support Vector Machine, Bagging and Boosting Ensembles, Random Forest, Gradient Boosting, XGBoost, and Decision tree architectures. In a benchmark study that compares 18 such algorithms, [56],Cedral Decision Tree emerged as the best technique in terms of prediction accuracy.Among six different machine learning algorithms (Random Forest, XG Boost, SVM, KNN, Logit, and K-means) that were tested for process outcome prediction, upon six data sets, based on majority voting of the results, XGBoost emerged as the best performing algorithm, however, not with a statistically significant differentiation from the performance of Random Forest and Logit algorithms [59]. To predict the remaining time, which is a regression problem [52] XGBoost performed best against 4 other ML algorithms. Verenich et al. [68] combined XGBoost (in the form of a classification algorithm) together with process flow analysis to predict the remaining time. Another ensemble approach that uses multiple ML algorithm is the stacked architecture with XGBoost and LightGBM in combination with a feature vector that considered inter-case resource dependencies [55]. To predict remaining time, random forest was used as a classification algorithm to determine the remaining process path, and then used it as an indirect mechanism to predict the remaining time of the process [64].

When used in machine learning algorithms, the event logs are required to be specially treated to be converted into meaningful features. Firstly, the process prefix traces (refer chapter 1) are often being bucketed (segregated into bins), to introduce homogeneousness of data that is fed into the machine learning model. Once that is done, the prediction is obtained by training multiple models, each trained for a specific bucket of prefix traces. The methodologies that are used for prefix bucketing include bucketing by the prefix length, clustering techniques, and using domain expertise [59, 68]. Once the process prefixes are bucketed, the process sequences need to be converted into features that capture the sequentiality of the process trace. Unlike when using tabular data, the sequential nature and order of event sequences need to be captured in event log data when being fed to machine learning models, to resemble the actual process. This is achieved by encoding the attributes of event sequences using an appropriate sequence encoding mechanism. Such mechanisms include Last State Encoding (*considering only the last state of the process prefix and its attributes*),Aggregation Encoding(*When the event attributes are fed as feature categories, with an aggregation of the number of occurrences of that attribute within the prefix trace and/or duration to complete the event is used as the feature value*) and Index Encoding (*representing each attribute of each $i^l h$ event as an individual feature, resulting in a very high feature space dimensionality*).

Sequence encoding would often result in loss of information in event logs, hence use of Deep Learning techniques have recently gained popularity in predictive process analytics. Deep learning has the inher-

ant ability to work with large feature space dimensionalities, and specifically Recurrent Neural Networks (RNN) by design are well catered towards handling sequential data, thus being suitable for event logs.

LSTM networks are popularly used in next event prediction [17, 18, 57], due to their ability of not only processing sequential data, but also the ability of remembering the dependencies of previously executed process events towards the final prediction. There are also modifications towards the vanilla LSTM architecture proposed in various studies. Generative Adversarial Networks (GAN) based models [58], The ensemble LSTM architectures to handle multiple event attributes [8] and Key-value-predict attention network (KVP) [23] are some of those modified architectures.Despite being originally proposed to handle image data, Convolutional Neural Networks (CNN) are also a popular choice in predictive process analytics domain, likely due to their ability to handle high dimensional feature sets [23, 42, 43]. When using CNN, the common approach used is to convert the event logs into an ordered array of event attributes, that resembles a 2D feature matrix, similar to a pixelized image. For categorical feature encoding in deep learning techniques, embedding [53], word-embedding and one-hot-encoding [57] are used commonly. Unlike the sequence encoding techniques that are required to be used in traditional machine learning approaches, the information loss of these techniques is minimal.

2.2 Explainable AI Techniques and Evaluation

This section attempts to understand i) the existing approaches available to interpret the Machine learning and Deep learning models; ii) Why we need explanations/ model explanation purposes and iii) evaluation techniques available for those explainability techniques. The model explanation techniques can be divided into four categories, based on the nature of the explanation and how the explanation is derived [20].

- *Black-box model explanation:* A surrogate model is used to approximate the black-box in terms of model performance, with comparable accuracy and fidelity, and gives a comprehensive global explanation. The techniques that falls in this category include
- *Black-box outcome explanation:* A surrogate model is used to approximate the black-box with good accuracy and fidelity and gives a local explanation for a given instance of the outcome, explaining how the model arrived at that particular decision. E.g.: Tree-PAN algorithm [14].
- *Black-box inspection:* An externally applied technique or a function that takes the outcome of the black-box and the data set to derive on a visualization that explains the behavior of the black-box. LIME [47, 66], which establishes a linear relationship between the inputs and a given output, SHAP [19], which assesses the influence of an input, based on a game theoretical principle, and Partial Dependence Plots [35] are all black-box inspection techniques.
- *Transparent model:* A machine learning model itself is (internally) interpretable (e.g., a decision tree, a linear regression model), and has a mechanism to provide a transparent explanation (e.g., extracted rules of a decision tree, feature weights of linear regression, or a logistic regression model). Explaining a model with the internal components of the model is also called intrinsic explanation. The concept for intrinsic explainability is also explored for not so transparent deep learning models as well, by using internal model properties to explain the models. E.g.: Attention weights [1, 40, 74] which try to explain a model using the attention weights that are computed using LSTM hidden states, Regression activation maps [25, 73] which uses the final layer (global average pooling layer) pixel attribution of the CNN to explain the models, Layer wise relevance propagation [70].

2.2.1 Explanations with a Purpose

Explaining a model is essentially understanding how the model makes the decision, either by decoding the computational logic, or by understanding the influential criteria that it uses to make a particular decision. This is often motivated by an end objective or a purpose, which could belong to one of the eight different broader explanation objectives [41].

1. Transparency - Understand how does the system work

2. Effectiveness - Help making good decisions
3. Trust - Improve the user confidence upon the system and its decisions
4. Persuasiveness - Convince users to try
5. Satisfaction - Increase the enjoyability of the system
6. Education - Help users to learn from the system
7. Efficiency - Help users to make fast decisions
8. Scrutability - Help users to identify if the system is broken, and makes decisions on an invalid basis
9. Debugging - Facilitate users to fix the system, if broken

Whilst purposes 1 - 7 focus on generating explanations to persuade the end users to trust the machine learning based predictive system, and justify the system, 8 and 9 focus on the system itself. It is crucial for a system to be robust, with a satisfactory level predictive performance. In a user oriented survey done to identify what people really need when they say 'explainable AI' [7], it became evident that Debugging Model and Identifying Bias became two of the top three reasons, that were ranked by 60 stake holders. Model explanations which helps achieving those purposes can be extracted at various stages of the model development [15].

Explanation of Underlying Data Also called pre-model explanations [9], are those that help to identify the nature and limitations of raw data, and relationship among the features. Lu et al. [32] in their evaluation of recent progress of predictive visual analytics, state that the visual explanations on the underlying data helps to identify the data errors and data losses conveniently at the stage of data preparation stage, and in the feature selection stage it helps efficient prioritization of features even without a prior domain knowledge.

Explanation of Model Performance Whilst it is common to state the model performance by overall performance metrics such as model accuracy, precision and recall, a deep dive of model performance helps to understand the model in detail. This deep dive can be done by dissecting the model performance further. E.g: By the target variable, By input data properties, By individual models (in an ensemble architecture).

Explanation of Model Decision Logic Explanation of model decision logic can be done by using either intrinsic properties of the model or by the use of post-hoc methods. Whilst the intrinsic interpretations generated for interpretable models such as decision trees and regression models do interpret the entire model, intrinsic interpretations of black-box models can only partially explain the model (e.g.: Attention weights, Regression Activation Maps, Layer wise relevance propagation). Intrinsic models are mostly dependent on internal model properties such as weight matrices. Post-hoc interpretation techniques such as LIME and SHAP, in contrast uses an externally applied techniques to make a guess on the how model makes a decision by evaluating the model inputs and outputs.

Granularity of Interpretation The most granular level model interpretation is the local interpretation. i.e. at the sample level. This is too granular to make a decision about the model's decision logic flaws that impacts model performance, however would be useful for a purpose of providing a justification of a particular decision to an end user. The highest level of interpretation is Global level, that is a summary of all the sample interpretations. According to [46], it is suggested that it is crucial to answer the questions 'why' and 'why not' at a global model level for the purpose of model inspection and improvement. However, it is not effective when the model complexity or underlying data complexity is quite high. Chan et al. [10] proposed an algorithm for cluster level summary of local interpretations. The clusters are generated by evaluating the similarity of the local interpretations. However, the clustering of interpretations can be also achieved heuristically. For an example, for the purpose of model inspection, if the model performance is too low for a particular target variable in a classification task, we can generate the cluster level interpretation, isolating the samples with that target variable being the ground truth (predictand).

What information needs to be extracted Once the level of interpretations that needs to be investigated is established, what information that needs to be extracted, and what summarizing technique to be used will be governed by explanation purpose. Model explanator types can be broadly categorized as *Feature Importance, Component Data (Data points with model outcome), Model Internals (Algorithmic representations), Explanation by Case, Contrastive, Counterfactual, Evidence (All information that supports a diagnosis), Prediction Certainty, Input Data, System Performance* [21, 51].

2.2.2 Evaluating Model Explanations

The effectiveness of a model explanation can be assessed by *Interpretability* – How easy it is to be understood by a human, *Accuracy* – How close is the accuracy between the surrogate model and black-box, *Fidelity*: How truthful the explanation is to the black-box's computation [75]. Whilst Interpretability is a common criteria for both post-hoc and intrinsic techniques, Fidelity is required to be assessed for post-hoc techniques, because the model and explanation technique are independent from each other, and Accurary is required to assessed for intrinsic techniques, because the model is changed.

There are two approaches that can be used to evaluate a model explanation given the different levels of concerns/interests and different end users [16].

Application- or Human-grounded evaluations. This approach tests the explanation based on human-centric experiments. When used the opinions of the domain experts, it is referred to as an Application grounded evaluation. Human-grounded evaluations are conducted using the opinions of lay people. In such an evaluation, both qualitative and quantitative metrics can be used to assess the explanation. Such qualitative metrics include Usefulness, Satisfaction, and Trust upon the explanation. The quantitative metrics can be the time consumed to understand the explanation, to the extent which the explanation is understood and the likelihood to deviate from the expected message communicated. In application- or human-grounded evaluations, it is crucial to have a basic understanding on the expectations of a human audience.

- *How humans collect evidences to reason*: Wang et al. [69] proposed a framework that elaborates on human reasoning practice based on various cognitive theories, and how XAI support that reasoning. According to the framework, humans are intrigued to understand a phenomenon, that is either unusual, or of specific interest. This is defined as the reasoning goal. Once a goal is established, humans make decisions by collecting evidences on a 'top-down', stemming from a hypothesis and validating with evidences (inductive) or 'bottom-up'(deductive) manner. Another approach is use of analogy to reason a similar instance. The reasoning process can be 'fast', that is, based on minimum evidences arriving at a conclusion based on the closest similar example, or 'slow', that is, considering all available factors and evidences.

- *What do humans expect as explanations*: It is suggested in social science theories that, humans expect constrastive explanations, where they can weigh the explanation for a certain decision, against an explanation for another contrastive or a similar decision. Also humans favour an explanation with only a handful of key reasons selected (often with a conformity bias), instead of an exhaustive reasoning, irrespective of the underlying statistical significance [37]. Further, it is observed that complex explanations (with a high number of rationales provided in the explanation), reduces the human satisfaction and increases the time spent in understanding the explanations [39]

Functionally-grounded evaluations. This approach uses computational metrics to evaluate the model explanation. Assessment of the fidelity of the explanation, The computational efficiency of generating explanation, and the explanation conciseness are used in functionally-grounded evaluations.

- *For surrogate model-based approaches (black-box explanation, black-box outcome explanation, or transparent black-box techniques)*: *Fidelity tests* check how many predictions of the black-box model did the surrogate model get right [5, 14, 26]. Further, the surrogate model size and computational efficiency, can be used as parameters of evaluating how concise the explanation is [75].

- *For feature attribution-based approaches (black-box inspection techniques)*: *Input Perturbation* [70] is performed by deleting/randomizing the highest scored inputs. *Sensitivity-n* [75] is an evaluation method to test the explanation of feature gradient-based explanations of neural networks.

2.3 Explaining Process Prediction Models

In an attempt of demystifying the reasoning logic of process predictions, researchers have focused on various post-hoc and intrinsic approaches of explaining those predictive models, and resulting predictions [46]. The importance of opening the black box lies in the need for humans to comprehend how the automated systems (which are often called black-boxes due to their computational complexity) make decisions, to ensure the validity and fairness of such decisions [50], especially in high stake decision making, which may involve a decision about a human life [48].

The interest upon explainability in predictive process models seem to be stemming mainly from the motivation towards the actionability [45] upon process predictions. Understanding the underlying problems in a process is one of the main use cases for explainable AI in process prediction domain. Some efforts of that use case include improvement of navigation process of an online job application website [35] and understanding the process instances that lead to expensive process path [36]. Another use case of XAI in predictive process analytics was using the explanations to improve the predictive performance of the model itself [47]. However, other than a handful of approaches, explainability in predictive process analytics is not geared towards achieving a particular purpose with model explanations, but ad-hoc efforts of trying to explain process prediction models.

In literature, both post-hoc and intrinsic approaches have been used to explain a predictive process models. Post-hoc methods try to explain a model by either using a simplified surrogate model to approximate the black-box model, or by using externally applied computation that tries to establish explainable relationships between inputs and outputs of the black-box. Intrinsic approaches mainly include the use of more interpretable and transparent models (e.g.: Decision Trees, Regression Models) to make the predictions [20]. However, the use of internal properties of a black-box (such as weight matrices) to partially explain the black-box is also considered as intrinsic explanation mechanism.

As the approach to this review, it was started with seven existing systematic literature review papers (including five reviews on PPA and two on XAI). Next, snowballing search (comprising a sequence of backward and forward searches) was conducted on these seven literature review papers. In parallel, a keyword search was conducted separately using 'Google Scholar' to capture other relevant and recent literature that were not amongst the above papers.

2.3.1 Post-hoc explanation of process prediction models

SHAP [33],is a method of explaining a black-box based on the Shapely value (in game theory) of a given input. SHAP is used to explain an LSTM-based outcome prediction model [19]. Rizzi et al. [47], Sindhagatta et al [54]. and Velmurugan et al. [66] used LIME to explain complex ML based process prediction models. LIME is a computational technique that establishes a linear relationship between inputs and outputs of the model to establish the influence of an input towards the prediction. Mehdiyev and Fettke have used surrogate decision trees [35] and partial dependence plots [36] to explain simple Feed Forward Neural Network (FFNN) based process predictions. One major drawback of post-hoc explanations is explanation fidelity (so that there is a requirement of ensuring fidelity via rigorous testing) [67]. Thus, post-hoc explanations may not be appropriate especially for purposes like inspecting process prediction models or diagnosing issues of underlying processes.

2.3.2 Intrinsic explanation of process prediction models

With the RNNs being popular choice in predictive process models, the attention weights, originally designed to improve the model performance are being used to obtain explanations[53, 1].Attention weights as a mechanism of explaining a LSTM [53] and Transformer networks [1] are used to derive local explanations for next event prediction, to identify the relative importance of input features/event attributes that

include the activity label, resource person who executed the activity for a given output. Layer wise Relevance Propagation (LRP) [70] is a technique of back propagating the relevance of each neuron towards the outputs (from the final layer to input layer). As the final explanation, LRP provides a feature attribution at single prediction level (local explanation). Another intrinsic explanation could be the use of hybrid approach of process modelling and Deep Learning to create a semi-transparent black box by Harl et al. [22]. They achieved this by implementing a Gated Graph Neural Network (GGNN) upon the derived process model with event logs. However, attention weights and other intrinsic explainability techniques that are used upon deep-learning models could only partially explain the model, because it only represents one computational element. Given the complexity of deep neural networks, whilst it is far-fetched to expect complete transparency the same way as simpler models, there is potential to improve the explainability of these models via innovation of more explainable deep learning architectures.

Chapter 3

Research Questions

In the domain of Predictive Process Analytics, it can be observed that the existing efforts on explaining 'black-box' process prediction models are conducted in an ad-hoc manner. Some of the studies are potentially concerned with an explicit explanation purpose, including model-debugging (relevant to data scientists) [47], process optimization (relevant to business analysts) [36], or outcome justification (relevant to customers) [35]. Further, the efforts on intrinsic explanation techniques for deep learning-based process prediction models seem to be limited to few approaches. The intrinsic approaches can be considered superior for deep learning-based models as they are truthful to the model and well aligned with addressing the purposes of model debugging and improvement. Once the explanation is generated, to cater an established explanation purpose, it is important to assess the effectiveness of the explanation, and test its functionality and comprehensibility [20]. Below are three research questions (RQs) identified from the literature review.

- **RQ1: What is the purpose of generating 'explanations', and what would be the different levels of 'explanations' in addressing that purpose?**

  A model explanation can be motivated by and serve a variety of purposes. The extraction of model explanation and the representation of explanation needs to be catered for an explanation purpose. Therefore, the first research question to address in this research is about identifying a purpose of generating explanation (e.g., for model inspection), and establishing the levels of (e.g., at various granularity) for model explanation given the identified purpose (and potential end users).

- **RQ2: What explainable techniques can be built upon the deep learning-based process prediction models to generate explanations (to address a given purpose)?**

  Due to the high computational complexity of a deep learning-based model, often a simple surrogate model is used as an explanation technique without concerning the internal structure of a prediction model, while intrinsic explanation techniques have only been able to partially explain such a model. However, an intrinsic explanation technique has the strengths in its ability of staying true to the prediction model. Thus the need arises to develop an intrinsic technique that can help explaining a deep learning-based process prediction model. To achieve that, it is important to consider the explainability of the model by design. The key design decisions of such a model include the choice of prefix bucketing technique and feature encoding method, and design of the internal model architecture.

- **RQ3: How to assess the effectiveness of model explanations developed with different explainable techniques (developed in addressing RQ2), given the purpose and levels of explanations (established in addressing RQ1)?**

  Evaluation of model explanations is still an emerging area of research. Whilst certain criteria have been proposed for consideration when evaluating model explanation (e.g., interpretability, accuracy and fidelity), the actual evaluation is often context-dependent, especially for interpretability. RQ1 focuses on identifying what would be the required level(s) of model explanation for a given purpose, which can be used to set such a context. Evaluation of explanations arrived at using the explainable techniques developed in RQ2 can be further broken down into two questions: i) what criteria can

be used to measure the effectiveness of a model explanation, and ii) what methods can be used to conduct evaluation.

Chapter 4

Methodology and Research Design

4.1 Research Objectives

The proposed research project intends to explore and develop novel intrinsic model explanation techniques for deep learning-based process prediction models. It further attempts to generate and align the explanations for a given purpose such as model inspection (oriented towards data scientists) and process inspection (oriented towards process analysts). Based on the research questions laid in Chapter 3, this research aims to achieve the following three objectives.

- **Objective 1:** Create a framework for generating purpose-driven explanations (addressing RQ1). It provides guidance on what explanation to extract from a process prediction model and how to present the model explanations to end users for a given purpose.
- **Objective 2:** Develop intrinsic explainable techniques for deep learning-based process prediction models (addressing RQ2). These techniques will be built on different deep learning architectures and apply different methods to support explainability to help derive insights on the strengths and weaknesses of various models.
- **Objective 3:** Develop an evaluation framework to assess the effectiveness of model explanations developed for a given purpose, with the proposed explainable techniques (addressing RQ3). The framework should specify what aspects of the model explanation will be evaluated, including the quantitative and qualitative metrics that can be used to measure those aspects, and a guidance (or methods) on how to conduct assessment using those metrics. Once the evaluation framework is established, it will be used to assess model explanations.

4.2 Research Methodology

Henver et al. [49] established a framework of 7 guidelines for an effective Design Science Research Project. The 7 guidelines are developed upon the principle that a Design Science Research (DSR) Project is essentially a problem-solving process and in the course of executing the project, proper knowledge and understanding will be acquired upon the underlying solution, as well as its solution.

In summary, the 7 guidelines propose that the end outcome of a Design science research project needs to be a viable IT artifact, that is the solution for a valid business problem, which is evaluated for its effectiveness as per the criteria that are set forth, and is a novel contribution to the research in terms of either the artifact, theory or the methodology, conducted with rigor and as a process of continuous search and knowledge discovery, and the results and outcomes are effectively communicated for both technology-oriented and business-oriented audiences.

The 7 guideline framework, in relevance to the proposed project, will be as follows:

- **Guideline 1: Design as an artifact –** The end outcomes of the proposed research project will be a.) A framework that guides the generation of model explanations geared towards a specific explanation purpose. b.) A methodology of extracting intrinsic model explanations from a deep learning-based process prediction model. Whilst the outcome (a) is an artifact which is conceptual in nature, the outcome (b) is a technical artifact.
- **Guideline 2: Problem Relevance –** The research project tries to address problem of non-harmony between the human decision-makers and machine decision-making systems, due to the lack of transparency of latter. And more specifically, it addresses the problems of absence of clear guidance of generating model explanations towards a specific purpose, and the insufficiency of intrinsic model explanation techniques for deep learning-based process prediction models, that will cater the specific purposes of model inspection and process inspection.
- **Guideline 3: Design Evaluation –** As a part of the research, the developed frameworks and techniques will be tested for the effectiveness via both user-based experiments and computational approaches.
- **Guideline 4: Research Contribution –** The proposed research project addresses the gaps identified in existing research via a Literature review, and expected to deliver novel contributions in the area of 'Explainability in Deep Learning-based process predictions' that is under-researched.
- **Guideline 5: Research Rigor –** The project will be carried out at the expected level of rigor, which will be ensured by publishing the findings in peer reviewed publications, and thorough evaluations.
- **Guideline 6: Design as a Search Process –** The research will be conducted as an iterative process of discovering/ conceptualizing and developing new approaches, evaluating and modifying or eliminating them until it reaches the desired research objectives.
- **Guideline 7: Communication of Research –** The outcomes of the research project will be presented to the research community as well as interested parties from the industry in the appropriate forms that can be research papers, online publications, and presentations.

4.3 Research Plan

The project will contain three main phases – development of the framework for purpose-driven explanations, development of intrinsic explainability techniques for deep learning-based process prediction models, and evaluation (RQ3). These phases may overlap along the research timeline.

A research plan can be defined in accordance with the framework proposed by Peffers et al. [44] for conducting Design Science Research. As per the framework, a DSR process may start with identification of an existing problem, definition of objectives of a solution, design and development of an artifact or a demonstration of an existing artifact, for a given context. At this stage of the project, the research objectives have been established to address the problems that have been identified. Thus, the following research plan is proposed to carry out this research.

Phase	DSRM Phase	Activities
Phase 1 (Development of the framework for purpose-driven explanations)	Define Objectives of Solution	Understanding the theoretical and experimental background on what are the purposes that are satisfied by model explanations. Identifying different aspects of a model explanations. Understanding how those aspects can be customised to develop an explanation which suits a given purpose.
	Design and Development	Derivation and validation of a methodology/ framework that guides developing 'purpose-driven' model explanations.
Phase 2 (Development of intrinsic explainability techniques for deep learning-based process prediction models)	Define Objectives of Solution	Identifying the key characteristics of an intrinsic model explanation. Investigating what kind of deep learning-based models suit process prediction. Understanding the existing intrinsic explainability techniques that are applied upon such models, and how appropriate are they for process predictions.
	Design and Development	Developing Deep learning-based Process Prediction Models to generate predictions. Development of intrinsic explainability techniques for those models, based on the background research carried out that suits the purposes of model inspection and process inspection.
	Demonstration	Application of developed explainability techniques upon an appropriate problem context, and demonstration of results.
Phase 3 (Evaluation)	Define Objectives	Identifying how the effectiveness of intrinsic model explanation could be measured. Identifying the criteria that measure the appropriateness of a model explanation for a given purpose.
	Design and Development	Develop an evaluation framework to assess the effectiveness of the model explanations developed with the proposed explainable techniques, for the purpose they are generated.
	Evaluation	Evaluation of demonstrated explainability techniques via evaluating the model explanations developed based on the evaluation framework that is developed
Continuous	Communication	Communication to the research community via publishing journal papers and participating in research conferences. Thesis write-up to communicate the overall contribution of the research

Table 4.1: Research plan of the proposed project based on the DSRM framework proposed by Pefferx et. al [44]

4.4 Resources and Funding Requirement

Hardware requirements: There will be no additional hardware requirements that are foreseen to conduct the research effectively, in addition to the hardware resources provided by QUT. If there would be a constraint on processing large sets of data, it is expected to use a cloud resource specialized for machine learning. E.g., Google Co-lab, AWS Sagemaker.

Software requirements: Python will be the primary coding language for the project due to the extended data manipulation capabilities and availability of a vast pool of relevant libraries. In addition, T-SQL will be used for data manipulation For development, open-source environments like Jupyter Notebook (with Anaconda distribution) or PyCharm will be used. Disco will be used for the initial data pre-processing stage of event logs, for process discovery. For the visualization of explanations, Microsoft Power BI (QUT licensed) version will be used.

Datasets: As the main form of data used for experiments, real-life event logs of process executions will be used.These will be sourced from publicly available sources. E.g., 4TU Centre for Research Data.

Evaluations: If the project requires human evaluations, it will be required to liaise with domain experts and business users who would interact with similar processes. This will be done after the necessary ethics approvals, and may require funding to incentivise the subjects.

4.5 Individual Contribution

Within our research group (`https://www.xami-lab.org/`), there are multiple research projects which explore the areas of process analytics and explainable AI. Some research projects aim at understanding the attributes of event log data that impact process predictions. In explainable AI, the on-going research focuses on evaluation of explanations, narrowing human-model gap with effective explanation representations and specific types of techniques for generating human-understandable explanations (e.g., counterfactuals). Figure 4.1 depicts an overall research scope for explainable predictive process analytics.

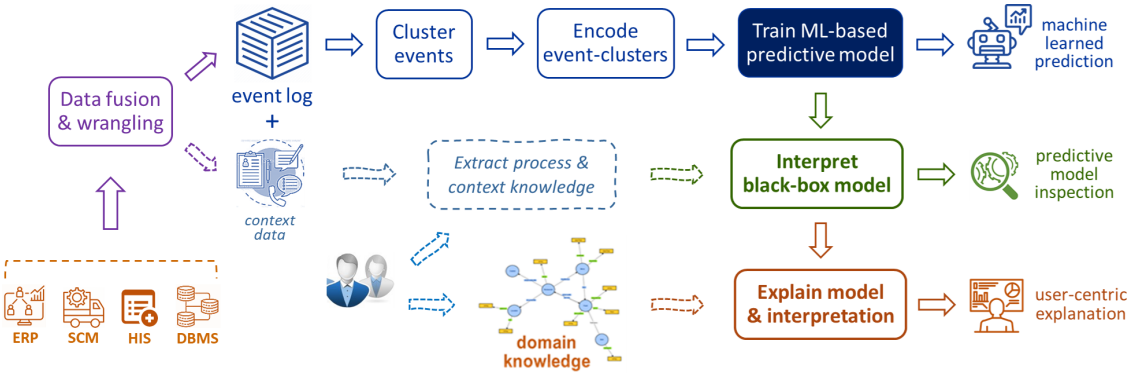


Figure 4.1: Research Scope of Explainable Predictive Process Analytics

To achieve the overall objectives of the research group, my research would contribute will following elements.

- Proposal of a framework for purpose-driven model explanations, as opposed to ad-hoc approach for generating model explanations.
- Introduction of new intrinsic model explanation techniques for deep learning-based process prediction models.

Chapter 5

Progress to Date

5.1 Purpose-driven Explanations

5.1.1 Five-phased methodology for generating purpose-driven explanation

To address the first research question of this research project (RQ1) established in chapter 3, a five-phased methodology is proposed, that guides generation of model explanations that would cater to a specific given purpose.

Figure 5.1: A five-phased methodology for generating purpose-driven explanations

This methodology is developed based on the concepts informed by various literature, inclusive of those included in chapter 2. The five phases of the methodology are as follows;

Establish Purpose: A model explanation could be used for a variety of end purposes. Such purposes may range from achieving the *transparency* of the decision, building *trust*, *persuading* the human stakeholders, achieving system *scrutability* (i.e identifying if the system is wrong), and *debugging* the system [41]. Establishment of the exact purpose for which we try to explain the model helps to determine the nature and complexity of the explanation [61].

Extract Explanation: It is suggested that model explanations can be either extracted by analysing the elements that go into the machine learning model *(pre-model explanations)* and/or by analysing the internal computational logic of the model *(post-model explanations)* [9]. The primary sort of explanation is the post-model explanation, as the focus of model explanation is to understand how the machine learning model makes the decision based on the feature set/ data points it is fed. This can be achieved either by using a post-hoc or an intrinsic technique (chapter 2). These primary (post-model) explanations can be supported by pre-model explanations, which could include the analysis of underlying data, problem context, domain knowledge and feature annotations.

Sort Explanations: The explanations that are extracted are often at a very granular level, or in a format that is not comprehensive enough. In order to develop meaningful interpretations out of the model explanations, we need to determine at which level we present those explanations *(explanation granularity)* and what information to be presented *(Insights)*.

The most granular level of model explanation is a *Local explanation*, which refers to the explanation of a single prediction. The highest granularity *Global explanation* refers to explaining the model holistically. Whilst, the Global explanation may help to determine why does a model behave in a given way, and overall which criteria it pays attention to, it may not be quite effective if the underlying data complexity is high [46]. However, a local explanation could be not sufficient when the purpose of explaining the model is concerned about the entire model as opposed to a single sample, such as model inspection and debugging. As a solution, a good middle ground proposed is *Cohort-level explanations* [10], that allows the aggregation of explanations for similar samples into subgroups either by using explicit criteria or a clustering technique. At each granularity that is suitable for the purpose, we can draw insights from the explanations. Such insights can be classified into 'symptoms', 'potential causes' and 'support evidences'. Symptoms are those elements/ observations that help to identify an issue in the prediction of interest. Potential causes can explain why the symptom is. This can take the form of feature attribution, decision rules or a counterfactual explanation [51]. Support evidences are helpful to further strengthen the explanation, which can take the form of feature annotations, domain knowledge or exploratory analysis of data.

Present Explanations: Presenting the extracted and sorted explanations is a crucial step to convert the technical explanations into a humanly understandable format. This can be achieved in a variety of modes including Logical/algorithmic representations, textual/natural language based representations, or visual [2, 30] presentations.

Customise Explanation: Explanations may also required to be customised either based on the domain or specific user group [21, 37]. How the information is presented, the complexity of explanations [39], and making explanations relevant the application domain [30, 31, 69] are main customization considerations.

5.1.2 Framework for Model Inspection with Explanations

Guided by the five-phased methodology of developing purpose-driven explanations, following framework is proposed. This framework defines an approach for developing model explanations for the purpose of model inspection in the application domain of predictive process analytics is developed.

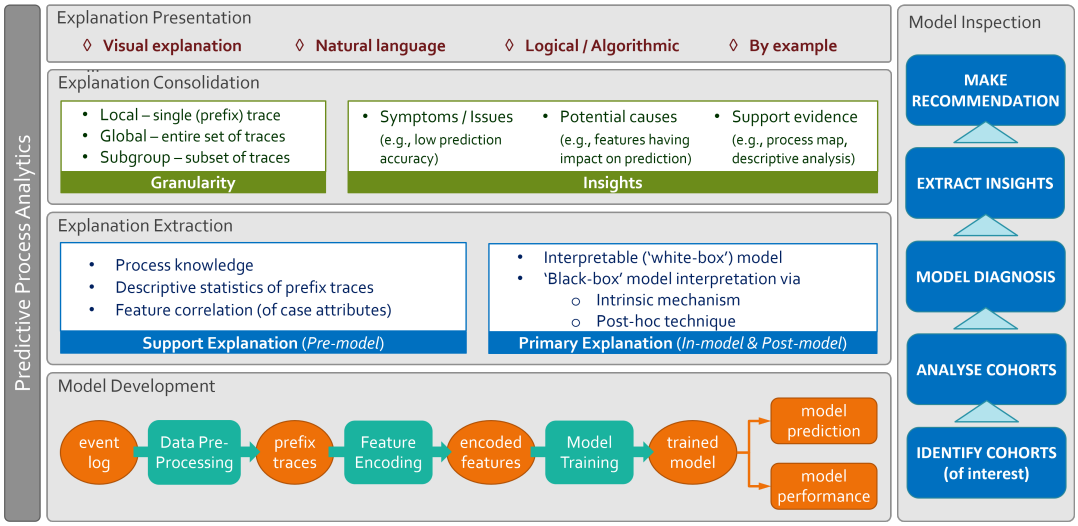


Figure 5.2: Framework for developing model explanations for model inspection

Model Development: This layer of the framework represents the starting point of the model explanation, which is the machine learning model itself. At each stage of the model development cycle [72], model explanations can be extracted [15]. In particular, the annotations and user notes available in a process event

log and the process domain knowledge can contribute to the model explanation with an equal importance as the the explanations that are extracted by opening up the predictive model.

Explanation Extraction To achieve the purpose of model debugging, we could extract two levels of model explanations.

- Primary explanations: The reasoning logic used by the model to make a particular decision is referred as the primary explanation. This can be extracted via a post-hoc or intrinsic explanation mechanism, when the model is not a white-box.

- Support explanations: Support explanations can either be symptoms or support evidences. Symptoms in the context of model inspection could be cohorts of poor predictive performance. Support evidences are those information that can make the claims made by the primary explanation more concrete (e.g.: descriptive statistics in event log, process domain knowledge)

Explanation Consolidation and Presentation This layer is to represent and guide the process of converting extracted model interpretations into meaningful explanations that would suit the purpose of model inspection

- Explanation Granularity: The granularity [10, 46] at which explanations will be presented. Whether it is for a single instance (local) or for a cohort of instances, or for the entire model is determined based on the expected purpose. For model inspection, it is rather important to understand why does the model behave in a certain way for either all or a large group of data. Hence, either global or cohort level granularity could be more appropriate.

- Insights: For the task of model inspection, symptoms that we will be after is often poor or too good predictive performance. Poor performance could be rationalized by examining the influential feature that lead to false predictions [47] or the issues in underlying data [11]. This can be supported with process knowledge and descriptive statistics of prefix-traces.

Model Inspection: The key criteria of assessment of a machine learning model are model performance, and model interpretability/ transparency [62]. For the purpose of model inspection, we focus on the aspect of model performance. Some approaches of model inspection/ debugging and model improvement with explanations include; analysing the explanations of model sub cohorts [29], understanding the features/ data points that leads to incorrect predictions [29, 47, 11], or correct predictions predictions [38]. Once the root causes of incorrect predictions are identified, they could be corrected by either removal of faulty features or data points and re-training the model.

Based on the foundation laid by the experimental studies on different approaches of using model explanations for model inspection/ debugging and improvement, we propose a 5-stepped approach on how a model could be debugged with the help of model explanations.

- Specify model cohorts (model partitioning based on inputs)

- Identify problematic cohorts that display poor predictive performance

- Use primary model explanations to identify which features/ properties of the inputs result in such poor performance and arrive at a primary diagnosis that is specific to the cohort

- Strengthen the primary diagnosis with support explanations

- Recommendation for fixing the identified issue

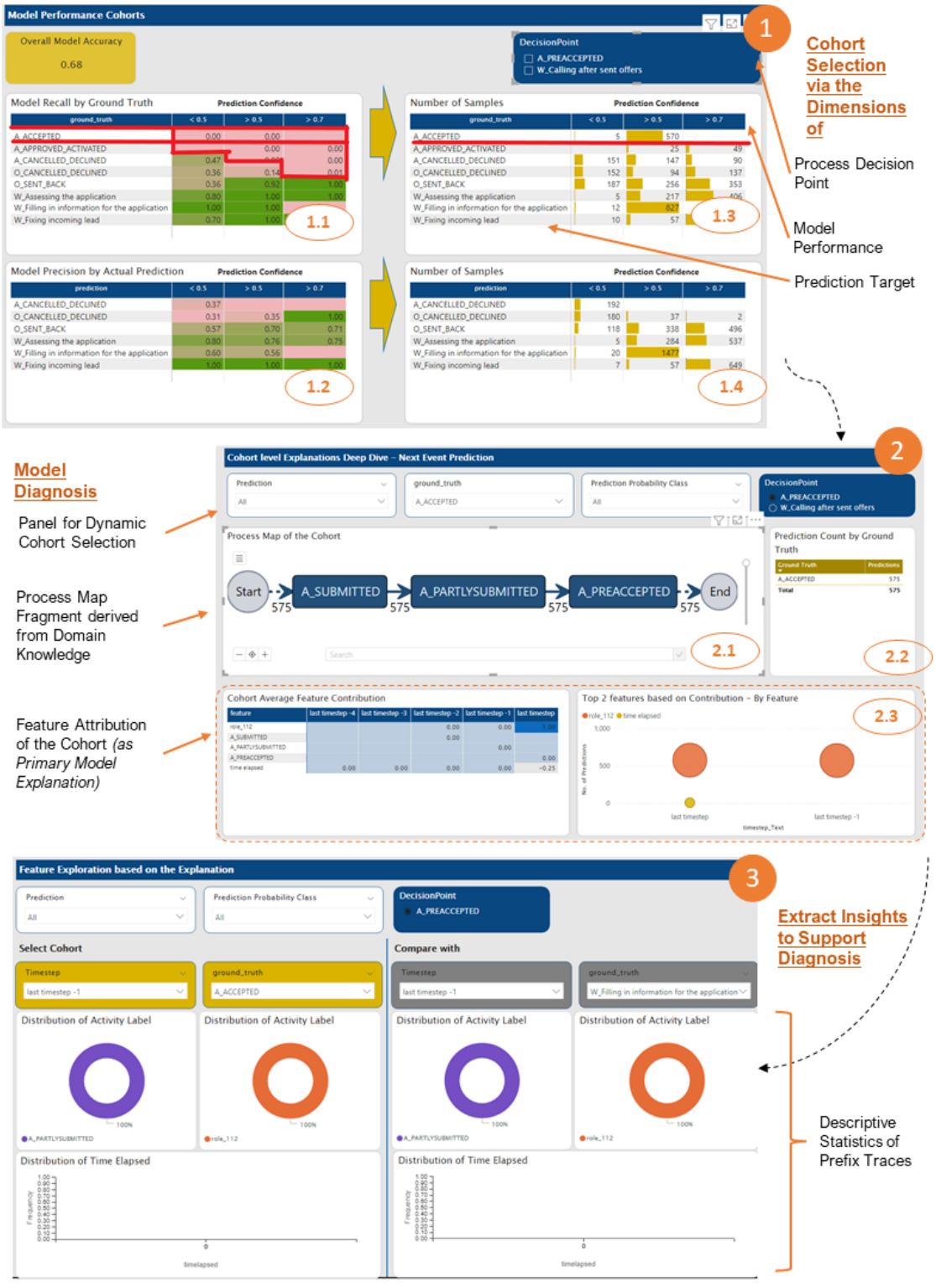


Figure 5.3: Visual Model Explanations for Model Inspection

5.1.3 Case Study: Explanations catered for inspection of a Process Prediction Model

To inspect a LSTM based next activity prediction model [71] with its relatively poor performance in certain prediction targets, following visual analytics platform based on model explanations is used (Figure 5.3).

The explanations from the model were extracted using the attention weights of the LSTM network. Attention weights are an intrinsic explanation extraction mechanism, and it was chosen as a more appropriate approach for the purpose of model inspection against possible post-hoc approaches. The reason being, the intrinsic approaches are inherently truthful to the model, hence will indicate exactly how the model makes decisions. The approach of extracting explanations will be detailed in the next section (i.e., Section 5.2 on 'Novel Attention-Based Model Explanation Mechanism').

Based on the proposed 5-stepped approach of model inspection, the visual analytics platform is organized. Frame 1 of Figure 5.3 is intended to identify problematic model cohorts. It is achieved by initially subdividing the model performance by specified cohorts. In this implementation, (a) process decision point at which the prediction was made, (b) Prediction (predicted value) were used as cohort defining criteria. Then, considering prediction recall and precision as the main performance indicators to measure the predictive performance of each cohort, the poor performing model cohorts are identified (sub frames 1.1 and 1.2). Sub frames 1.3 and 1.4 indicate how many predictions fall into each prediction cohort.

Frame 2 of Figure 5.3 allows the user to select the model cohort they want to inspect using the selection pane on the top. Once the cohort is selected, sub frame 2.3 depict the primary model explanation (Attention weights-based) pertaining to that cohort. Left hand side of the sub frame 2.3 averages out the feature weights of all the samples of the cohort selected, and shows in average what is the contribution of different features towards the decision. The right hand side shows the top 2 most influential features (based on attention weight) contributing to the prediction, considering all the samples. It can be observed that the explanations are depicted along the time step of the process prefix trace, which is a unique requirement in process prediction model explanations. Sub frame 2.1 and 2.2 are support evidences (pre-model explanations) that would help to better understand the primary explanation. Sub frame 2.1 depicts the partial process map of the selected sub cohort. The process map carries the domain knowledge about the underlying process, and shows how the process would be generally executed. Sub frame 2.2 shows the breakdown of the samples of the cohort by prediction target (or ground truth). This helps to evaluate the level of misclassification of a given prediction.

Frame 3 of Figure 5.3 provides a distribution of process prefix features across the samples of the selected cohort. This helps to further support any conclusions that was arrived using the explanations in Frame 2.

Paper Accepted:

Bemali Wickramanayake, Chun Ouyang, Catarina Moreira, and Yue Xu. Generating Purpose-Driven Explanations: The Case of Process Predictive Model Inspection. Accepted by the *34th International Conference on Advanced Information Systems Engineering (CAiSE 2022) Forum*, 6 - 10 June, 2022.

5.2 Novel Attention-Based Model Explanation

Attention weights of LSTM and Transformer networks are traditionally used as a mechanism of improving the model performance. Attention weights have an ability to point out the most influential steps of an input sequence that is passed through a LSTM or a Transformer network. This property can be used to partially explain a LSTM/ Transformer based deep learning model [1, 53].

Explainability of a model starts from the design of the model itself, when an intrinsic model explanation technique such as attention weights are used. Having this in mind, we propose a novel design of interpretable model for next activity prediction of a given process trace, inspired by the work by Camargo et. al [8].

We explain the model using a single prefix trace denoted by $P = e_1, ...e_l$, where e is a single event of the process prefix trace, and l is the prefix length (i.e. the number of events). An event e_i can be represented as

a tuple of event attributes $(a_{e_i}, r_{e_i}, t_{e_i})$, or (a_i, r_i, t_i) as a simplified notation. a_i denotes the activity that was executed in that event. r_i denotes the resource person/ role of the resource person carried out the activity. t_i denotes the time attribute, which is often either the start or the end time stamp of the event (or both). Although an event could carry other event attributes in addition to a, r, t, however in this model the interest is limited to these three attributes.

The model takes a prefix trace $([(a_1, r_1, t_1), ..., (a_l, r_l, t_l)])$ as the input and predict the next activity (a_{l+1}) for the prefix trace.

The model takes each attribute of a prefix trace as an independent sequence of inputs, to predict the next activity. Each attribute goes through a specialised LSTM layer independently, and the combined effect of the outputs of each individual LSTM is used to make the final prediction. Hence we give it the name 'specialised attention-based model'.

To explain the model, we use two types of attention weights. α - the event attention and β - the attribute attention. The attribute attention is computed for each attribute individually, and denoted by $\beta_a, \beta_r, \beta_t$.

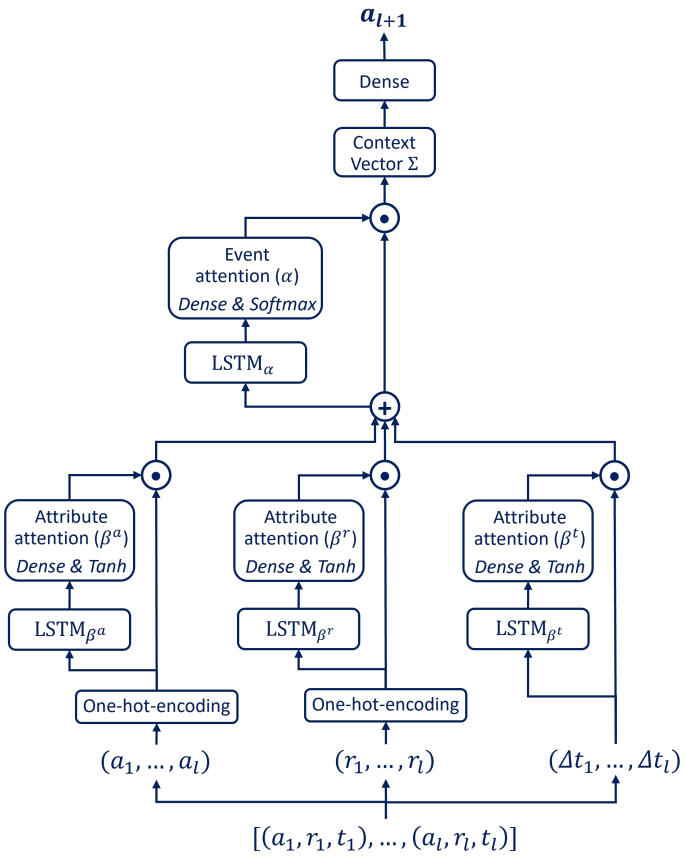


Figure 5.4: Specialised Attention-Based Model Architecture

The idea behind using two types of attention weight vectors for model explanation is, it is vital not only to understand which events do affect the final prediction, but also to which extent the different event attributes of a, r, t influence the prediction.

5.2.1 model configuration and predicting next activity

Figure 5.4 depicts the architecture of the proposed model. Out of the three input features, a, r, t, the two categorical features of a, r will be one-hot-encoded before being fed into the model. The one-hot encoded feature vectors are denoted by v_a^{ohe} and v_4^{ohe} . To represent the time attribute, we compute the feature 'time elapsed', which is the time difference (in hours) between the first time stamp of the prefix trace (start of the

prefix trace), and the time stamp of the given event. The time elapsed feature will be denoted as v_t.

$$(a_1,...,a_l) \rightarrow v_a^{ohe} \in \{0,1\}^{l\times|A|}$$
$$(r_1,...,r_l) \rightarrow v_r^{ohe} \in \{0,1\}^{l\times|R|}$$
$$v_t = (\Delta t_1,...,\Delta t_l)$$

Attribute attention β is computed by passing each feature vector through an individual BiLSTM layer — $LSTM_{\beta_a}$, $LSTM_{\beta_r}$, and $LSTM_{\beta_t}$ — The BiLSTMs generate hidden state vectors $g^a \in \mathbb{R}^{l\times|A|}$, $g^r \in \mathbb{R}^{l\times|R|}$, and $g^t \in \mathbb{R}^l$, respectively.

$$g^a \leftarrow LSTM_\beta^a(v_a^{ohe})$$
$$g^r \leftarrow LSTM_\beta^r(v_r^{ohe})$$
$$g^t \leftarrow LSTM_\beta^t(v_t)$$

Using the hidden state vectors, the attribute attention vectors for each feature category $\beta^a \in \mathbb{R}^{l\times|A|}$, $\beta^r \in \mathbb{R}^{l\times|R|}$, and $\beta^t \in \mathbb{R}^l$ are computed by passing them individually through three *tanh* dense layers as follows.

$$\beta_{ij}^a = tanh(W_\beta^a g_{ij}^a + b_\beta^a)$$

where $i \in \{1,...,l\}$, $j \in \{1,...,|A|\}$, and W_β^a and b_β^a are the trained weight matrix and bias of the *tanh* dense layer used to compute β^a, respectively.

$$\beta_{ij}^r = tanh(W_\beta^r g_{ij}^r + b_\beta^r)$$

where $i \in \{1,...,l\}$, $j \in \{1,...,|R|\}$, and W_β^r and b_β^r are the trained weight matrix and bias of the *tanh* dense layer used to compute β^r, respectively.

$$\beta_i^t = tanh(W_\beta^t g_i^t + b_\beta^t)$$

where $i \in \{1,...,l\}$, and W_β^t and b_β^t are the trained weight matrix and bias of the *tanh* dense layer used to compute β^t, respectively.

Next the computed attention vectors are multiplied with the relevant feature vectors,to arrive at intermediate vectors which represents the 'influence' of each feature at feature specific level.

$$v_a^{inf} = v_a^{ohe} \odot \beta^a$$
$$v_r^{inf} = v_r^{ohe} \odot \beta^r$$
$$v_t^{inf} = v_t \odot \beta^t$$

Three influence vectors are then concatenated and passed through a single BiLSTM layer to obtain the event attention α which considers the influence of all three features simultaneously as follows.

$$v_{a,r,t}^{inf} = v_a^{inf} \cup v_r^{inf} \cup v_t^{inf}, \text{ where } v_{a,r,t}^{inf} \in \mathbb{R}^{l\times(|A|+|R|+1)}$$

To compute **event attention** α, we use the hidden state vecor generated by $LSTM_\alpha$, which is denoted by $h \in \mathbb{R}^{l\times(|A|+|R|+1}$), taking the concatenated feature level influence vector $v_{a,r,t}^{inf}$ as the input.

$$h \leftarrow LSTM_\alpha(v_{a,r,t}^{inf})$$

Event attention $\alpha \in \mathbb{R}^l$ is then computed via an immediate *softmax* dense layer as follows.

$$\alpha_i = softmax(W_\alpha^T h_{ij} + b_\alpha)$$

where $i \in \{1,...,l\}$, $j \in \{1,...,|A|+|R|+1\}$, and W_α and b_α are the trained weight matrix and bias of the *softmax* dense layer, respectively.

To combine the impact of event and attribute attention,event attention α is element-wise multiplied with the concatenated feature influence vector $v_{a,r,t}^{inf}$, and aggregated (by addition) over the prefix length to obtain a context vector c, which computes the cumulative effect of event and attribute level influence over the entire prefix. This context vector then goes through a dense layer to calculate the final prediction — the next activity a_{l+1} of the input prefix trace.

$$c = \sum^l \alpha \odot v_{a,r,t}^{inf}$$

5.2.2 Extracting Model Explanation

The model explanations extracted from this model are in the form of feature weights. Feature weights indicate the relative importance of a given feature, compared to the rest of other features that were fed into the model. Due to the model architecture, where the categorical features of 'activity' and 'role' were fed into the model as one-hot-encoded features, each feature value of these two features are acting as its own independent feature. Hence, from this point on the feature values of 'activity' and 'role' referred to as 'features' and the event attributes of 'activity' and 'role' are referred to as 'feature categories'. The feature weights are extracted using the attention weights of both α and β attention layers.

Computation of feature weights with attention weights: Initially, for each prediction sample, the relevant local explanation was extracted. To obtain the combined impact upon the features for a given prediction sample, first the event attention vector α was multiplied with the attribute attention vectorβ, element wise. This results in a 2D feature weight matrix, for each event, and each feature of the input process prefix. Due to orthogonality of the features among the same feature category facilitated by one-hot-encoding, the non-zero values of the i^th row of the feature weight matrix represents the feature weights of the event $e_i = (a_i, r_i, t_i)$, that was fed into the model.

$$\begin{gathered}\text{feature weight of } a_i = \text{non-zero value of } \alpha_i \odot \beta_i^a \\ \text{feature weight of } r_i = \text{non-zero value of } \alpha_i \odot \beta_i^r \\ \text{feature weight of } t_i = \text{non-zero value of } \alpha_i \odot \beta_i^t\end{gathered}$$

Identifying the most influential events (time steps): For the representation purposes to identify which events influenced the prediction decision more, we use the α attention weights. To identify the n most influential events of the prefix (with a length l), where $n \leq l$, we use the n highest attentions of the from the α attention layer, for the given prefix.

Deriving Global Explanations: Instead of deriving global explanations for the entire model, the interest was to identify what features would most commonly affect for a given prediction. To achieve this, we averaged out the locally obtained feature weights, for a given prediction.

5.2.3 Evaluation

The proposed model architecture was evaluated for both model performance, and model explainability. To evaluate the model performance, the model performance was compared against a number of experimental benchmarks [8, 17, 27, 34, 43, 57, 60], and to evaluate the explainability, explanations are compared against a baseline architecture [53], designed to explain model explanations with attention weights. To further validate the explanations, they were compared against the process domain knowledge [3].

Dataset: The primary dataset used for the experimental evaluation is an event log that belongs to a loan approval process in a Dutch financial institution *BPIC 2012* [6]. This dataset was made publicly available for the annual business process challenge at 8th international workshop on Business Process Intelligence (2012). The motivation behind using this specific dataset is, the familiarity of a loan application process to a wider audience including non-process domain experts and the availability of credible benchmarks for necessary evaluations.

The dataset contains event log data that belongs to three sub processes. Application sub processes (indicated by the A at the beginning of the relevant activity labels), Offer sub process (O), and Work item sub process(W). However, all three sub processes are closely interlinked to create the loan application process hence, we consider all three sub processes in our experiments.

To further validate the consistency of the model performance, a secondary dataset *BPIC 2017* [65] is used, which is a more recent event log of the same (yet further improved) loan application process as *BPIC 2012*.

An event log contains both dynamic (event specific) and static (case/ trace specific) features. The dynamic features change over the execution of the process, and static features stay constant.

Event Log	Num. cases	Num. activities	Num. event	Avg. case length	Max. case length	Avg. case duration	Max. case duration	Variants
BPI 2012	13087	36	262200	20.04	175	8.62 days	137.22 days	4366
BPI 2012 Complete	13087	23	164506	12.57	96	8.61 days	91.46 days	4336
BPI 2012 W Complete	9658	6	72413	7.5	74	11.4 days	91.04 days	2263
BPI 2012 O	5015	7	31244	6.23	30	17.18 days	89.55 days	168
BPI 2012 A	13087	10	60849	4.65	8	8.08 days	91.46 days	17

Table 5.1: Data profiles of event logs used in experiments.

Feature Category	BPIC 2012
Single Process Trace Identification	Case_ID
Dynamic Features	
Activity	Concept_Name
Role (Performer)	Resource
Timestamp	Complete_Timestamp
Lifecycle Transition	Completed
Static Features	None

Table 5.2: Event log features considered for the experiment.

Evaluation of model performance

The objective of model performance evaluation is to ascertain that the proposed model architecture yields in a reasonable predictive performance despite the explainability. Intrinsic model explanation techniques, similar to the proposed technique often affects the model architecture. Thus, it needs to be ensured that the model explainability does not come at the cost of model performance. The model performance is measured by the prediction accuracy.

	BPIC 2012 Complete	**BPIC 2012 A**	**BPIC 2012 O**	**BPIC 2012 W Complete**
Pasquadibisceglie et al. [43]	74.55	71.47	77.51	66.14
Tax el al. [57]	79.39	77.75	81.22	67.80
Camargo et al. [8]	79.22	78.92	85.13	65.19
Hinkka el al. [24]	79.76	79.27	85.51	67.24
Khan et al. [27]	75.50	75.62	84.48	75.91
Evermann et al. [17]	63.37	74.44	79.20	65.38
Di Mauro et al. [34]	78.72	78.09	81.52	65.01
Theis et al. (w/o attributes) [60]	73.10	66.23	81.52	76.97
Theis et al. (w/ attributes) [60]	65.21	65.12	73.56	72.52
Proposed Model (Specialised attention-based)	**77.74**	**74.86**	**81.60**	**83.75**

Table 5.3: Accuracy comparison in % Percentage w.r.t. benchmark results

To ensure the competitive performance of proposed specialised attention-based mode architecture against the explainable model used as a baseline [53], to compare the explainability , predictive performance of the two models were compared using the traditional performance metrics to accuracy, precision, recall and F1-Score, using both *BPIC 2012* and *BPIC 2017* datasets.

	Accuracy		Precision		Recall		F1-score	
	Baseline [53]	Specialised	Baseline [53]	Specialised	Baseline [53]	Specialised	Baseline [53]	Specialised
BPIC 2012 Complete	0.79	0.78	0.77	0.75	0.79	0.78	0.74	0.73
BPIC 2012 A	0.75	0.75	0.67	0.74	0.75	0.75	0.69	0.70
BPIC 2012 O	0.82	0.82	0.82	0.81	0.82	0.82	0.80	0.80
BPIC 2012 W Complete	0.84	0.84	0.85	0.85	0.84	0.84	0.84	0.83
BPIC 2017	0.83	0.82	0.83	0.82	0.83	0.82	0.82	0.81

Table 5.4: The performance of the *Baseline model* and *specialised attention-based model*.

Evaluation of model explainability

The primary two criteria of measuring the effectiveness of model explainability are the interpretability and fidelity of explanations [20]. Interpretability means the extent to which humans can comprehend and understand the explanation. Fidelity refers to the truthfulness/faithfulness of the explanation to the model. Intrinsic explanations are naturally faithful to the model, as they are extracted from the properties of the model itself. Thus, the evaluation of the model explanations will be limited to the interpretability aspect of it.

In this evaluation, the explanations analysed are pertaining to the predictions of next activity at process decision points. A decision point of a process is identified by an activity in the process which leads to multiple next events. The two decision points in the loan application process 'A_PREACCEPTED' and 'W_Calling after sent offers' are considered in this analysis. 'A_PREACCEPTED'(Figure 5.5) is a decision point that occurs early in the loan application process, and 'W_Calling after sent offers'(Figure 5.6) occurs in the middle of the process.

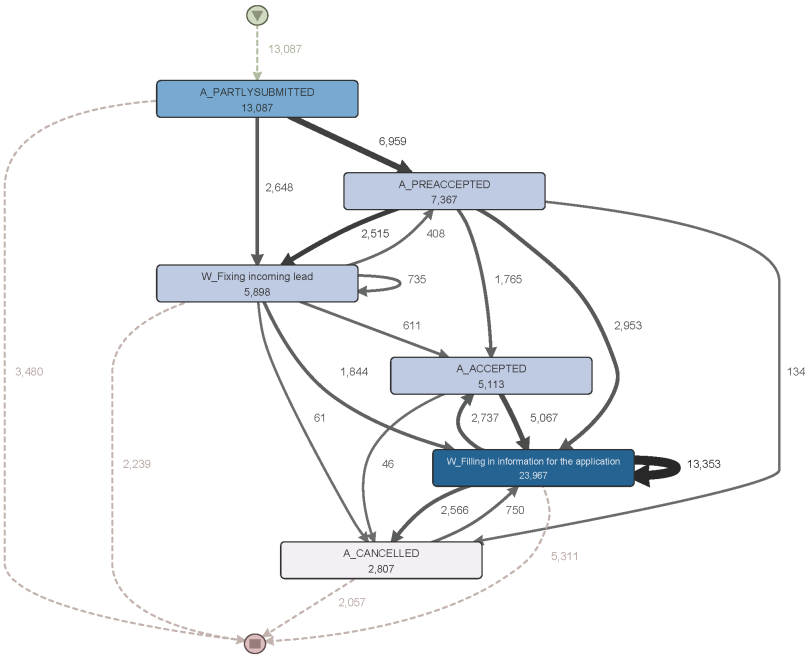


Figure 5.5: A partial process map for decision point A_PREACCEPTED.

Model explanations at 'A_PREACCEPTED': The model performance is relatively poor at this decision point in both specialised attention-based model (proposed model architecture) and baseline model [53]. Possible next activities at this decision point are 'W_Fixing incoming lead', 'A_Accepted' and 'W_Filling in information for the application', with a fair balance among the prediction targets. However, both models are only able to predict either 'W_Fixing incoming lead' or 'W_Filling in information for the application' targets.

As per the global explanation of the prediction target of 'W_Filling in information for the application' (Figure 5.7), it can be observed that both the models point out to the 'role_112' feature of the 'role' feature category at the last executed event to be the most influential feature towards the decision. In lay language, this means that, there is a very high bearing of the person who executed the last activity of the prefix trace (i.e the activity 'A_PREACCEPTED') being 'role_112' being, towards the next activity to be 'W_Filling in information for the application'. The local explanations for the same prediction, for both true (Figure 5.8)

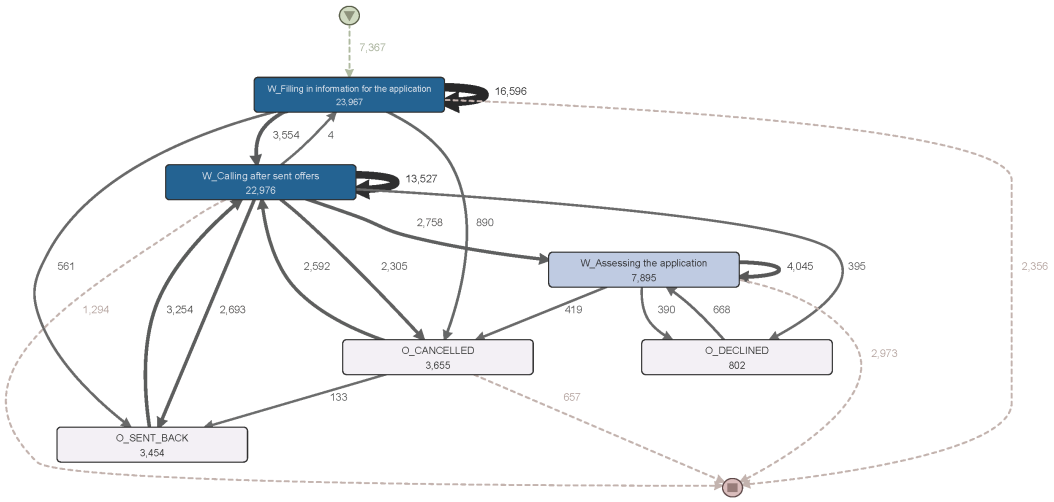
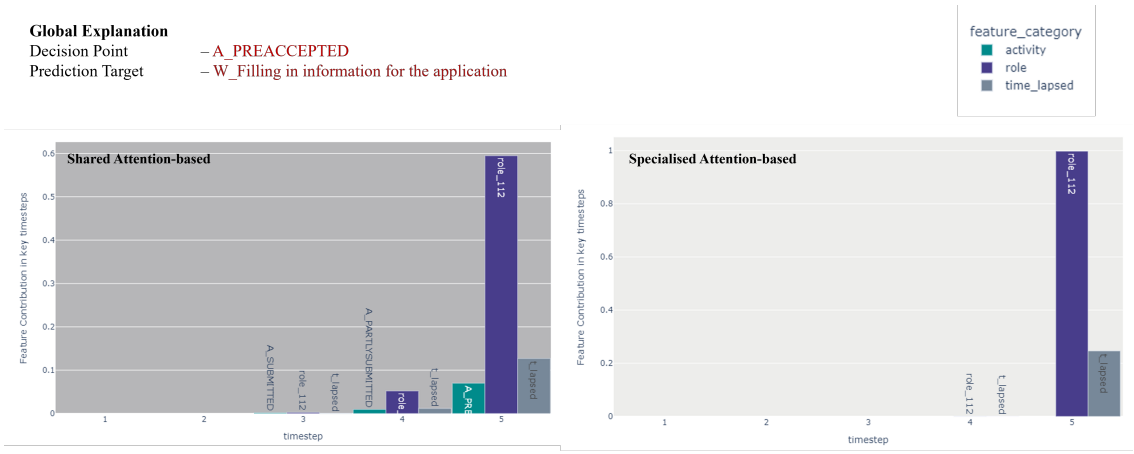


Figure 5.6: A partial process map for decision point W_Call for Sent Offers.



Accuracy	0.70	Accuracy	0.70
Target Recall	1.00	Target Recall	1.00
Target Precision	0.56	Target Precision	0.56
F1-Score	0.72	F1-Score	0.72

Figure 5.7: Global Explanation for 'W_Filling in information for the application' target at 'A_PREACCEPTED' decision point

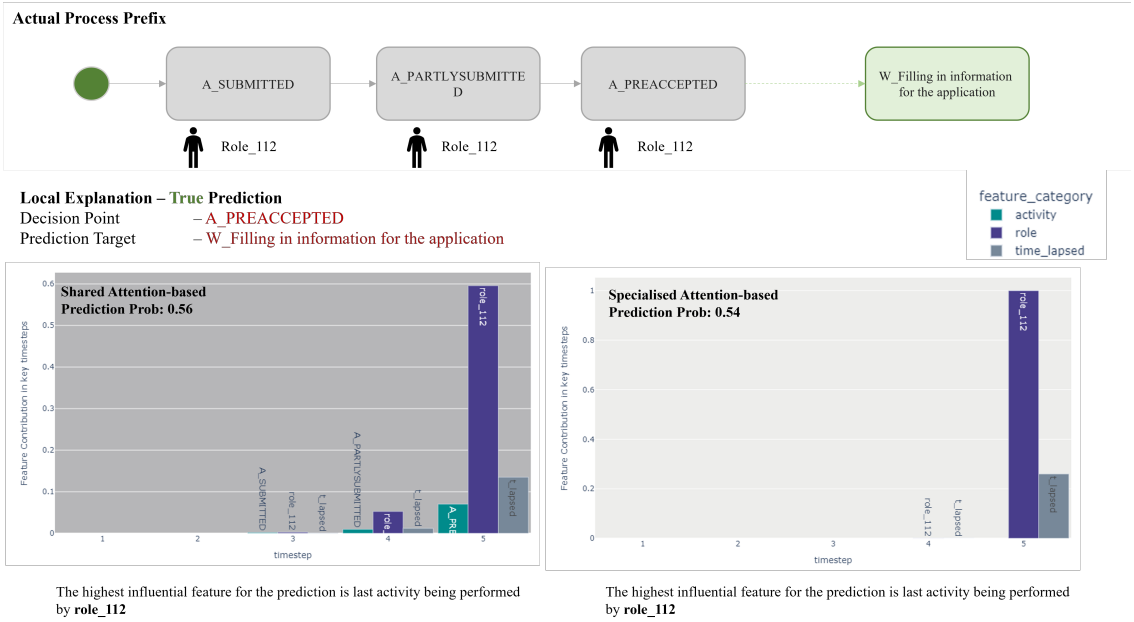


Figure 5.8: Local Explanation for a **True** prediction for 'W_Filling in information for the application' target at 'A_PREACCEPTED' decision point

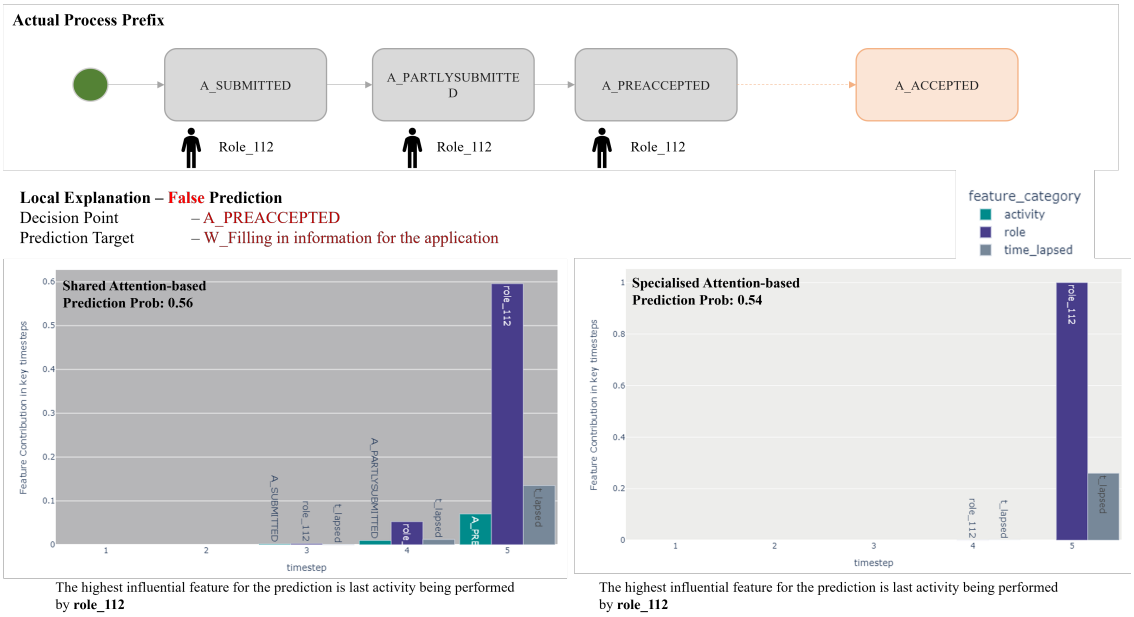
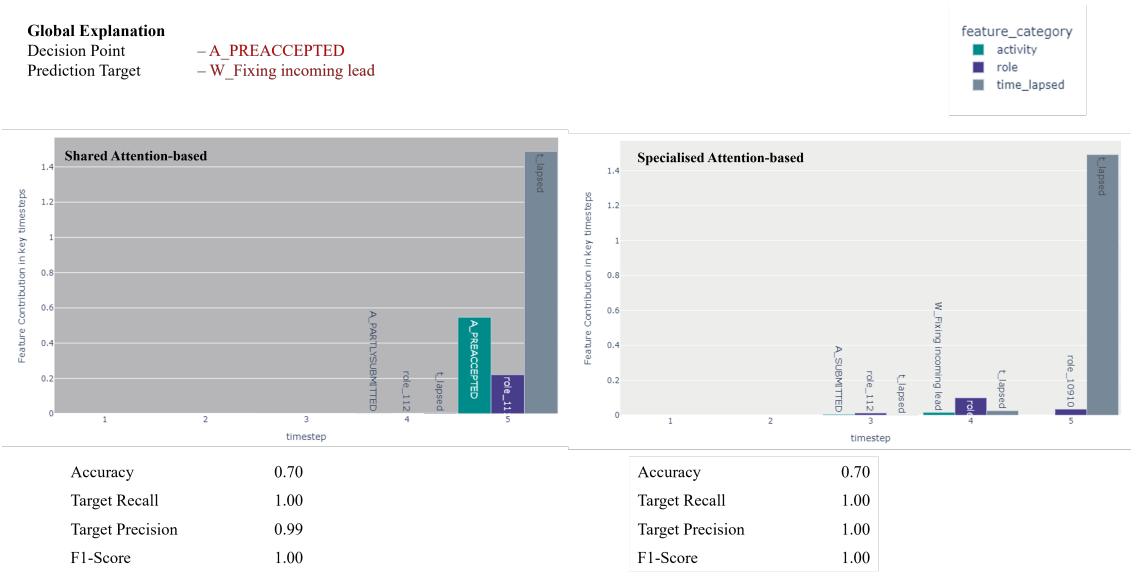


Figure 5.9: Local Explanation for a **False** prediction for 'W_Filling in information for the application' target at 'A_PREACCEPTED' decision point

and false (Figure 5.9) predictions remain consistent with the global explanation. However, the over dependency upon this one single feature could explain the low precision of this prediction target, as well as low prediction confidence, where almost all the predictions being made with a probability that does not exceed 0.55.



Accuracy	0.70
Target Recall	1.00
Target Precision	0.99
F1-Score	1.00

Accuracy	0.70
Target Recall	1.00
Target Precision	1.00
F1-Score	1.00

Figure 5.10: Global Explanation for 'W_Fixing incoming lead' target at 'A_PREACCEPTED' decision point

Figure 5.10 depicts the global explanation for the other prediction target at this decision point 'W_Fixing incoming lead'. Again, on this explanation both proposed and baseline models agree with each other, and indicates the most influential feature to be 'time lapsed' at the last activity of the prefix trace.

Model explanations at 'W_Calling after sent offers': This decision point refers to the followup activity done by the bank, once it sends a loan offer to the loan applicant. The action of sending a loan offer is denoted by the activity 'O_SENT'. Once the 'O_SENT' activity is performed, the bank will periodically follow up with the offer, until it receives an acknowledgement (denoted by the activity 'O_SENT_BACK'). If the bank does not receive an acknowledgement, it will then cancel ('O_CANCELLED_DECLINED') the application process [3].

As per the global explanation for the prediction 'O_SENT_BACK' (Figure 5.11), we can observe that the two models use different features to arrive at the decision. The baseline model focuses on the last 3 occurred 'W_Calling after sent offers' activities in the prefix trace, and associated 'role' features as more important to make the decision. The proposed model points out that the last 'W_Calling after sent offers' activities,and the two 'O_SENT' activities (and associated 'time lapsed' feature at each of those events) are more influential. This explanation made by the proposed model aligns much better with the domain expert explanation [3], compared to the baseline model.

However, the in the global explanation for the prediction 'O_CANCELLED_DECLINED', both model explanations agree with each other for some extent 5.12. Both the models consider the time lapsed features that are associated with 'W_Calling after sent offer' activities to be very influential towards the prediction. This means the model considers the time that was spent after each 'W_Calling after sent offer' activity determines if the prediction should be 'O_CANCELLED_DECLINED'. This aligns very well with the explanation by the process domain experts.

Accuracy	0.66	Accuracy	0.65
Target Recall	0.86	Target Recall	0.82
Target Precision	0.69	Target Precision	0.69
F1-Score	0.76	F1-Score	0.75

Figure 5.11: Global Explanation for 'O_SENT_BACK' target at 'W_Calling for sent offers' decision point

Accuracy	0.66	Accuracy	0.65
Target Recall	0.11	Target Recall	0.18
Target Precision	0.31	Target Precision	0.32
F1-Score	0.16	F1-Score	0.23

Figure 5.12: Global Explanation for 'O_CANCELLED_DECLINED' target at 'W_Calling for sent offers' decision point

Paper Revision under Review:

Bemali Wickramanayake, Zhipeng He, Chun Ouyang, Catarina Moreira, Yue Xu, and Renuka Sindhgatta. Building Interpretable Models for Business Process Prediction using Shared and Specialised Attention Mechanisms. 2nd revision submitted to *Knowledge Based Systems*, Elsevier.

Chapter 6

Future Work and Research Timeline

To achieve the research objectives stated for this project, the future work that needs to be carried out at each phase are as follows.

6.1 Phase One: Purpose-Driven Explanations

Currently the initial framework for the purpose-driven explanations has been proposed for the purpose of *model inspection*, which partially completes this phase. We plan to extend this to address an additional purpose *process inspection*. The existing framework is currently demonstrated upon model explanations for next activity prediction using an event log of a loan application process. The framework needs further validation with a different process prediction model and different event logs.

The initial demonstration of model inspection with explanation resulted in a diagnosis for the sub-par model performance for some of the prediction targets. This diagnosis requires to be validated by implementing the recommendations suggested after the diagnosis, and retraining the model to improve the model performance. This will objectively confirm the validity of the proposed framework of 'purpose-driven explanations for model inspection'.

6.2 Phase Two: Intrinsic Explainability for Deep Learning-based Process Predictions

This phase is also initiated with introducing specialised attention-based mechanism to explain a LSTM-based next activity prediction model. The future plans for this phase includes;

- Exploration of other intrinsic techniques that are used to explain deep learning models
- Adapt and demonstrate those techniques if they would be suitable for process prediction applications
- Customise and introduce novelty to those techniques with further developments including those in the model architecture, and explanation extraction mechanism, so that they generate explanations that are appropriate for process prediction models
- Validating the techniques with multiple event log datasets, and prediction problems

6.3 Phase Three: Evaluation

The demonstrations of phases one and two will required to be evaluated with appropriate evaluation criteria. The explainability techniques developed in phase two will be presented as final explanations based on the framework developed in phase one. At phase three, development of appropriate evaluation framework to evaluate those final explanations will be the primary task. Once that is done, actual evaluations will need

to be conducted to assess mainly the interpretability aspect of the explanations. Thus, at this stage, it will be required to carry out human-grounded evaluations.

6.4 Publication Strategy

The proposed framework for purpose-driven explanations (for model inspection) **section 5.1, Chapter 5** is currently accepted to the International Conference of Advanced Information Systems Engineering (CAiSE) 2022. The novel attention-based model explanation mechanism **section 5.2, Chapter 5** is currently under review for a special issue in the journal Knowledge Based Systems, and the archived version is available here [71]. Outside these two submissions, the future work of this projects is considered to for publications that include leading conferences and journals.

Amongst the conferences in interest are ACM SIGKDD International Conference on Knowledge Discovery Data Mining (h-index 104), IEEE International Conference on Data Mining (h-index 54) which focus on artificial intelligence and data mining and Int. Conference on Business Process Management (impact factor 3.3), a leading conference in business process management. The journals that are aimed for future work in this research include Knowledge and Information Systems (h-index 46) and Journal of Big Data (h-index 42).

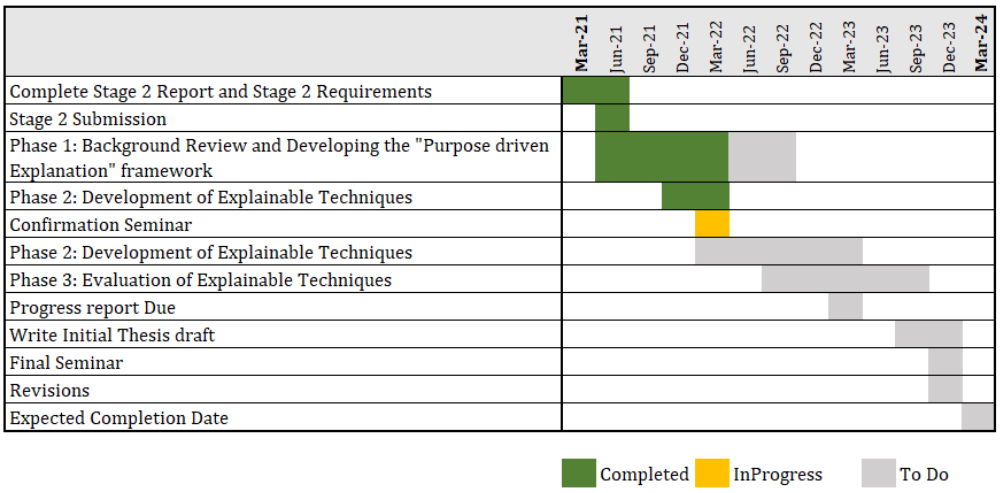


Figure 6.1: Research Timeline

Bibliography

[1] Prerna Agarwal, Daivik Swarup, Sushruth Prasannakumar, Sampath Dechu, and Monika Gupta. Unsupervised contextual state representation for improved business process models. In *Business Process Management Workshops*, pages 142–154. Springer International Publishing, 2020. 5, 8, 20

[2] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, jun 2020. 17

[3] Arjel D. Bautista, Lalit Wangikar, and Syed M. Kumail Akbar. Process mining-driven optimization of a consumer loan approvals process – the BPIC 2012 challenge case study. In *International Conference on Business Process Management*, volume 132 of *Lecture Notes in Business Information Processing*, pages 219–220. Springer, 2012. 23, 28

[4] Gaël Bernard and Periklis Andritsos. Accurate and transparent path prediction using process mining. In Tatjana Welzer, Johann Eder, Vili Podgorelec, and Aida Kamišalić Latifić, editors, *Advances in Databases and Information Systems*, pages 235–250, Cham, 2019. Springer International Publishing. 1, 3

[5] Olcay Boz. Extracting decision trees from trained neural networks. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '02*. ACM Press, 2002. 7

[6] BPI Challenge 2012. Event log of a loan application process, 2012. 23

[7] Andrea Brennen. What do people really want when they say they want "explainable AI?" we asked 60 stakeholders. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. ACM, apr 2020. 6

[8] Manuel Camargo, Marlon Dumas, and Oscar González Rojas. Learning accurate LSTM models of business processes. In *International Conference on Business Process Management*, volume 11675 of *Lecture Notes in Computer Science*, pages 286–302. Springer, 2019. 5, 20, 23, 24

[9] Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832, jul 2019. 6, 16

[10] Gromit Yeuk-Yin Chan, Enrico Bertini, Luis Gustavo Nonato, Brian Barr, and Cláudio T. Silva. Melody: Generating and visualizing machine learning model summary to understand data and classifiers together. *CoRR*, abs/2007.10614, 2020. 6, 17, 18

[11] Ching-Ju Chen, Ling-Wei Chen, Chun-Hao Yang, Ya-Yu Huang, and Yueh-Min Huang. Improving CNN-based pest recognition with a post-hoc explanation of XAI. *Soft Computing (In Review)*, aug 2021. 18

[12] European Commission. General data protection regulation. *arxiv: 1612.08468*, 2016. 1

[13] Breuker D, Matzner M, Delfmann F, and Becker J. Comprehensible predictive models for business processes. *MIS Quarterly*, 2016. 1, 3

[14] Tanusree De, Prasenjit Giri, Ahmeduvesh Mevawala, Ramyasri Nemani, and Arati Deo. Explainable AI: A hybrid approach to generate human-interpretable explanation for deep learning prediction. *Procedia Computer Science*, 168:40–48, 2020. 5, 7

[15] Shipi Dhanorkar, Christine T. Wolf, Kun Qian, Anbang Xu, Lucian Popa, and Yunyao Li. Who needs to know what, when?: Broadening the explainable AI (XAI) design space by looking at explanations across the AI lifecycle. In *Designing Interactive Systems Conference*. ACM, jun 2021. 6, 17

[16] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv*, 2017. 7

[17] Joerg Evermann, Jana-Rebecca Rehse, and Peter Fettke. A deep learning approach for predicting process behaviour at runtime. In *Business Process Management Workshops*, pages 327–338. Springer International Publishing, 2017. 5, 23, 24

[18] Joerg Evermann, Jana-Rebecca Rehse, and Peter Fettke. Predicting process behaviour using deep learning. *Decision Support Systems*, 100:129–140, aug 2017. 5

[19] Riccardo Galanti, Bernat Coma-Puig, Massimiliano de Leoni, Josep Carmona, and Nicolo Navarin. Explainable predictive process monitoring. In *2020 2nd International Conference on Process Mining (ICPM)*. IEEE, oct 2020. 5, 8

[20] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5):1–42, jan 2019. 1, 5, 8, 10, 25

[21] Mark Hall, Daniel Harborne, Richard Tomsett, Vedran Galetic, Santiago Quintana-Amate, Alistair Nottle, and Alun Preece. A systematic method to understand requirements for explainable ai (xai) systems, 2020. 7, 17

[22] Maximilian Harl, Sven Weinzierl, Mathias Stierle, and Martin Matzner. Explainable predictive business process monitoring using gated graph neural networks. *Journal of Decision Systems*, pages 1–16, jun 2020. 9

[23] Kai Heinrich, Patrick Zschech, Christian Janiesch, and Markus 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*, 143:113494, apr 2021. 5

[24] Markku Hinkka, Teemu Lehto, and Keijo Heljanko. Exploiting event log event attributes in RNN based prediction. In *Data-Driven Process Discovery and Analysis*, volume 379 of *Lecture Notes in Business Information Processing*, pages 67–85. Springer, 2019. 24

[25] Giacomo Iadarola, Fabio Martinelli, Francesco Mercaldo, and Antonella Santone. Towards an interpretable deep learning model for mobile malware detection and family identification. *Computers & Security*, 105:102198, jun 2021. 5

[26] Ulf Johansson and Lars Niklasson. Evolving decision trees using oracle guides. In *2009 IEEE Symposium on Computational Intelligence and Data Mining*. IEEE, mar 2009. 7

[27] Asjad Khan, Hung Le, Kien Do, Truyen Tran, Aditya Ghose, Hoa Dam, and Renuka Sindhgatta. Memory-augmented neural networks for predictive process analytics, 2018. 23, 24

[28] Wolfgang Kratsch, Jonas Manderscheid, Maximilian Röglinger, and Johannes Seyfried. Machine learning in business process monitoring: A comparison of deep learning and classical approaches used for outcome prediction. *Business & Information Systems Engineering*, apr 2020. 1

[29] Sanjay Krishnan and Eugene Wu. Palm: Machine learning explanations for iterative debugging. In *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics*, HILDA'17, New York, NY, USA, 2017. Association for Computing Machinery. 18

[30] Bum Chul Kwon, Min-Je Choi, Joanne Taery Kim, Edward Choi, Young Bin Kim, Soonwook Kwon, Jimeng Sun, and Jaegul Choo. Retainvis: Visual analytics with interpretable and interactive recurrent neural networks on electronic medical records. *IEEE Trans. Vis. Comput. Graph.*, 25(1):299–309, 2019. 17

[31] Hyunkwang Lee, Sehyo Yune, Mohammad Mansouri, Myeongchan Kim, Shahein H. Tajmir, Claude E. Guerrier, Sarah A. Ebert, Stuart R. Pomerantz, Javier M. Romero, Shahmir Kamalian, Ramon G. Gonzalez, Michael H. Lev, and Synho Do. An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets. *Nature Biomedical Engineering*, 3(3):173–182, dec 2018. 17

[32] Junhua Lu, Wei Chen, Yuxin Ma, Junming Ke, Zongzhuang Li, Fan Zhang, and Ross Maciejewski. Recent progress and trends in predictive visual analytics. *Frontiers of Computer Science*, 11(2):192–207, oct 2016. 6

[33] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st Conference on Advances in Neural Information Processing Systems (NIPS)*, 2017. 8

[34] Nicola Di Mauro, Annalisa Appice, and Teresa M. A. Basile. Activity prediction of business process instances with inception CNN models. In *Lecture Notes in Computer Science*, pages 348–361. Springer International Publishing, 2019. 23, 24

[35] Nijat Mehdiyev and Peter Fettke. Prescriptive process analytics with deep learning and explainable artificial intelligence. In *28th European Conference on Information Systems*. An Online AIS Conference, 2020. 5, 8, 10

[36] Nijat Mehdiyev and Peter Fettke. Explainable artificial intelligence for process mining: A general overview and application of a novel local explanation approach for predictive process monitoring. In *Studies in Computational Intelligence*, pages 1–28. Springer International Publishing, 2021. 8, 10

[37] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, feb 2019. 7, 17

[38] Vineel Nagisetty, Laura Graves, Joseph Scott, and Vijay Ganesh. xai-gan: Enhancing generative adversarial networks via explainable ai systems. *CoRR*, 2020. 18

[39] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation. *CoRR*, 2018. 7, 17

[40] Dominic A. Neu, Johannes Lahann, and Peter Fettke. A systematic literature review on state-of-the-art deep learning methods for process prediction. *Artificial Intelligence Review*, mar 2021. 3, 5

[41] Ingrid Nunes and Dietmar Jannach. A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction*, 27(3-5):393–444, oct 2017. 5, 16

[42] Gyunam Park and Minseok Song. Predicting performances in business processes using deep neural networks. *Decision Support Systems*, 129:113191, feb 2020. 5

[43] Vincenzo Pasquadibisceglie, Annalisa Appice, Giovanna Castellano, and Donato Malerba. Using convolutional neural networks for predictive process analytics. In *2019 International Conference on Process Mining (ICPM)*. IEEE, jun 2019. 5, 23, 24

[44] Ken Peffers, Tuure Tuunanen, Marcus Rothenberger, and S. Chatterjee. A design science research methodology for information systems research. *Journal of Management Information Systems*, 24:45–77, 01 2007. 13, 14

[45] Jana-Rebecca Rehse, Nijat Mehdiyev, and Peter Fettke. Towards explainable process predictions for industry 4.0 in the DFKI-smart-lego-factory. *KI - Künstliche Intelligenz*, 33(2):181–187, apr 2019. 8

[46] Mireia Ribera and Àgata Lapedriza. Can we do better explanations? a proposal of user-centered explainable ai. In *IUI Workshops*, 2019. 6, 8, 17, 18

[47] Williams Rizzi, Chiara Di Francescomarino, and Fabrizio Maria Maggi. Explainability in predictive process monitoring: When understanding helps improving. In *Business Process Management Forum*, pages 141–158. Springer International Publishing, 2020. 5, 8, 10, 18

[48] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019. 8

[49] Henver A.and March S.and Park J.and Ram S. Design science in information systems research. *Management Information Systems Quarterly*, 28:75–, 2004. 12

[50] Muller K. Samek W., Wiegand T. Explainable arificial intelligence: Understanding, vizualizing and interpreting deep learning models. *Discoveries, Special Issue*, 2017. 8

[51] Tjeerd A.J. Schoonderwoerd, Wiard Jorritsma, Mark A. Neerincx, and Karel van den Bosch. Human-centered XAI: Developing design patterns for explanations of clinical decision support systems. *International Journal of Human-Computer Studies*, 154:102684, oct 2021. 7, 17

[52] Arik Senderovich, Chiara Di Francescomarino, Chiara Ghidini, Kerwin Jorbina, and Fabrizio Maria Maggi. Intra and inter-case features in predictive process monitoring: A tale of two dimensions. In *Lecture Notes in Computer Science*, pages 306–323. Springer International Publishing. 4

[53] Renuka Sindhgatta, Catarina Moreira, Chun Ouyang, and Alistair Barros. Exploring interpretable predictive models for business processes. In *Lecture Notes in Computer Science*, pages 257–272. Springer International Publishing, 2020. 5, 8, 20, 23, 24, 25

[54] Renuka Sindhgatta, Chun Ouyang, and Catarina Moreira. Exploring interpretability for predictive process analytics. In *Service-Oriented Computing*, pages 439–447. Springer International Publishing, 2020. 8

[55] Xiaoxiao Sun, Wenjie Hou, Yuke Ying, and Dongjin Yu. Remaining time prediction of business processes based on multilayer machine learning. In *2020 IEEE International Conference on Web Services (ICWS)*. IEEE, oct 2020. 4

[56] Bayu Adhi Tama and Marco Comuzzi. An empirical comparison of classification techniques for next event prediction using business process event logs. *Expert Systems with Applications*, 129:233–245, sep 2019. 4

[57] Niek Tax, Ilya Verenich, Marcello La Rosa, and Marlon Dumas. Predictive business process monitoring with LSTM neural networks. In *Advanced Information Systems Engineering*, pages 477–492. Springer International Publishing, 2017. 5, 23, 24

[58] Farbod Taymouri, Marcello La Rosa, Sarah Erfani, Zahra Dasht Bozorgi, and Ilya Verenich. Predictive business process monitoring via generative adversarial nets: The case of next event prediction. In *Lecture Notes in Computer Science*, pages 237–256. Springer International Publishing, 2020. 5

[59] Irene Teinemaa, Marlon Dumas, Marcello La Rosa, and Fabrizio Maria Maggi. Outcome-oriented predictive process monitoring. *ACM Transactions on Knowledge Discovery from Data*, 13(2):1–57, jun 2019. 1, 3, 4

[60] Julian Theis and Houshang Darabi. Decay replay mining to predict next process events. *IEEE Access*, 7:119787–119803, 2019. 23, 24

[61] Richard Tomsett, Dave Braines, Dan Harborne, Alun Preece, and Supriyo Chakraborty. Interpretable to whom? a role-based model for analyzing interpretable machine learning systems. *CoRR*, June 2018. 16

[62] Shailesh Tripathi, David Muhr, Manuel Brunner, Herbert Jodlbauer, Matthias Dehmer, and Frank Emmert-Streib. Ensuring the robustness and reliability of data-driven knowledge discovery models in production and manufacturing. *Frontiers in Artificial Intelligence*, 4:22, 2021. 18

[63] W.M.P. van der Aalst, H.A. Reijers, A.J.M.M. Weijters, B.F. van Dongen, A.K. Alves de Medeiros, M. Song, and H.M.W. Verbeek. Business process mining: An industrial application. *Information Systems*, 32(5):713–732, jul 2007. 1

[64] Sjoerd van der Spoel, Maurice van Keulen, and Chintan Amrit. Process prediction in noisy data sets: A case study in a dutch hospital. In *Lecture Notes in Business Information Processing*, pages 60–83. Springer Berlin Heidelberg, 2013. 4

[65] Boudewijn van Dongen. Bpi challenge 2017, 2017. 23

[66] Mythreyi Velmurugan, Chun Ouyang, Catarina Moreira, and Renuka Sindhgatta. Evaluating explainable methods for predictive process analytics: A functionally-grounded approach. volume abs/2012.04218. 2020. 5, 8

[67] Mythreyi Velmurugan, Chun Ouyang, Catarina Moreira, and Renuka Sindhgatta. Developing a fidelity evaluation approach for interpretable machine learning, 2021. 8

[68] Ilya Verenich, Marlon Dumas, Marcello La Rosa, Fabrizio Maria Maggi, and Irene Teinemaa. Survey and cross-benchmark comparison of remaining time prediction methods in business process monitoring. *ACM Transactions on Intelligent Systems and Technology*, 10(4):1–34, aug 2019. 1, 3, 4

[69] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, may 2019. 7, 17

[70] Sven Weinzierl, Sandra Zilker, Jens Brunk, Kate Revoredo, Martin Matzner, and Jörg Becker. XNAP: Making LSTM-based next activity predictions explainable by using LRP. In *Business Process Management Workshops*, pages 129–141. Springer International Publishing, 2020. 1, 5, 8, 9

[71] Bemali Wickramanayake, Zhipeng He, Chun Ouyang, Catarina Moreira, Yue Xu, and Renuka Sindhgatta. Building interpretable models for business process prediction using shared and specialised attention mechanisms. *arXiv e-prints*, pages arXiv–2109, 2021. 20, 31

[72] Rüdiger Wirth and Jochen Hipp. Crisp-dm: towards a standard process modell for data mining. 2000. 17

[73] Aleksandra Wolanin, Gonzalo Mateo-García, Gustau Camps-Valls, Luis Gómez-Chova, Michele Meroni, Gregory Duveiller, You Liangzhi, and Luis Guanter. Estimating and understanding crop yields with explainable deep learning in the indian wheat belt. *Environmental Research Letters*, 15(2):024019, feb 2020. 5

[74] Qinghan Xue and Mooi Choo Chuah. Explainable deep learning based medical diagnostic system. *Smart Health*, 13:100068, aug 2019. 5

[75] Jianlong Zhou, Amir H. Gandomi, Fang Chen, and Andreas Holzinger. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5):593, mar 2021. 7, 8

Appendix A

Course Requirements

A.1 Research Integrity Online (RIO)

The Research Integrity Online (RIO) course was completed on 22nd March 2021 and a copy of the certificate is included in this submission.

Figure A.1: RIO Certificate

A.2 Coursework

- **IFN001 - Advanced Information Research Skills (AIRS)**. AIRS final course material was submitted DATE and received a passing grade on 18th June 2021.
- **INN700 - Introduction to Research**. Completed on 9th July 2021
- **INN701 - Advanced Research Topics**. Partially Completed on 18th November 2021. Final module will be taken in Semester 1-2022

A.3 Ethical Clearance

I am preparing an ethics application or request for variation to existing ethics clearance for submission after Confirmation.

Ethics application(s) will be prepared for the use of industry event logs and for evaluation of research outcomes with industry partners. The current design depends on public logs, and we will review this arrangement if and when industry partners become involved during the project.

A.4 Intellectual Property

For specific industry engagement that I (may) become a part of, I'll sign the required industry IP assignment agreement. For the rest of my Ph.D. research, I do not need to sign an IP assignment agreement.

We will publish conference papers and develop open-source software. We will review this arrangement if and when industry partners become involved during the project.

A.5 Health and Safety

I do not need to complete Health and Safety training.

This project does not involve biological, microbiological, biomedical and biochemical material.

A.6 Collaborative Agreement

I do not require a Collaborative Agreement.

A.7 Thesis Type

Traditional by Monograph (Chapters).

A.8 Plagiarism Check

I have attached the screen shot of the plagiarism report from the iThenticate tool in Figure A.2. This was done excluding word level matches from the sources in consideration.

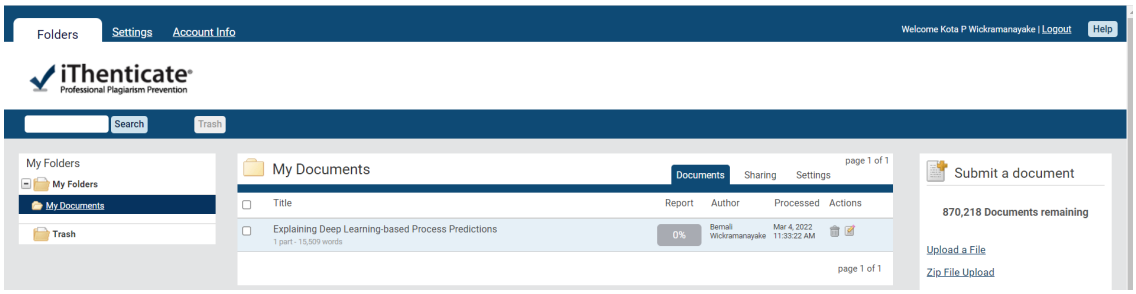


Figure A.2: Plagiarism Check