Improving Interpretability of Ensemble Surrogate Models for Clinical Prediction via Process Mining and Trace Clustering

Ehsan Baratnezhad

PhD student, Computer Science Department, Dalhousie University, NS, Canada

Supervisors: Dr. Syed Sibte Raza Abidi, William Van Woensel

Fall 2025

# Abstract

Table of Contents

[Abstract 2](#_Toc202354447)

[Chapter 1 Introduction 4](#_Toc202354448)

[1.1 Background 5](#_Toc202354449)

[1.2 Research Objectives 5](#_Toc202354450)

[1.3 Thesis Structure 5](#_Toc202354451)

[Chapter 2 Literature Review 6](#_Toc202354452)

[2.1 Theoretical Background 9](#_Toc202354453)

[2.2 Related Work 9](#_Toc202354454)

[Chapter 3 Methodology 10](#_Toc202354455)

[Chapter 4 Experiments and Evaluation 14](#_Toc202354456)

[Chapter 5 Conclusion and Future Work 20](#_Toc202354457)

[5.1 Discussion 24](#_Toc202354458)

[References 25](#_Toc202354459)

[Appendix A 26](#_Toc202354460)

# Chapter 1 Introduction

1.1 Motivation

Introduce the importance of interpretability in AI models, especially in clinical domains like kidney transplantation.

Highlight the limitations of black-box models and the need for explainable AI (XAI).

1.2 Problem Statement

Describe the challenge: multiple surrogate models lead to fragmented decision paths, making explanations incoherent.

Outline the difficulty in summarizing decision logic from ensembles of surrogates.

1.3 Research Objectives

Present the core aims of the research: enhancing interpretability through process mining, clustering decision paths, and identifying meaningful clinical patterns.

1.4 Research Questions

List the guiding questions such as:

* 1. Can common decision path patterns be discovered among surrogate models?
  2. How can process mining help visualize these patterns?
  3. What clinical insights can be derived from frequent and rare variants?

1.5 Methodological Approach

Overview of the approach: extracting decision paths → event log creation → trace clustering → process model discovery.

Justify the use of process mining for XAI enhancement.

1.6 Contributions

Summarize novel aspects such as:

* 1. Using process mining in reverse to aid XAI.
  2. Proposing a pipeline for structuring unstructured decision paths.
  3. Evaluating interpretability through clustering and process discovery metrics.

1.7 Thesis Outline

Provide a brief overview of the remaining chapters (Methods, Results, Discussion, Conclusion).

Context & Motivation

Importance of interpretability in clinical AI; challenge of black-box models in healthcare.

Problem Statement

Surrogate models produce too many unique decision paths; lack of coherent explanation.

Research Goals

Use process mining to extract interpretable patterns.

Analyze and cluster decision paths for simplification.

Contributions: List contributions like novel event log transformation, trace clustering for surrogate explanation, evaluation framework, etc.

Use

Sections from your paper’s introduction and abstract.

## 1.1 Background

## 1.2 Research Objectives

## 1.3 Thesis Structure

# Chapter 2 Literature Review

2.1 Foundations of Explainable Artificial Intelligence (XAI)

Define XAI and clarify distinctions:

Explainability vs. Interpretability vs. Comprehensibility

Local vs. Global Explanations

Ante-hoc vs. Post-hoc Methods

Relevance of XAI in clinical decision-making and trust-building

2.2 Surrogate Modeling for Post-hoc Explainability

Explain the surrogate modeling paradigm (e.g., decision trees, rule sets)

Model-agnostic vs. model-specific surrogates

Review of key techniques: LIME, SHAP, TREPAN, BETA, MUSE

Limitations of current surrogate-based explanations in complex models

2.3 Process Mining in Healthcare

Overview of process mining concepts (event logs, traces, variants)

Prior work applying process mining to clinical pathways

Challenges in applying process mining to decision paths:

Unstructured logic

Lack of temporal meaning

Need for clustering due to high variability

2.4 Clustering Techniques for Variant Simplification

Trace clustering in process mining (e.g., TF-IDF vectorization, spectral clustering, k-means)

Benefits of clustering to improve model interpretability

Evaluation metrics used (e.g., Silhouette score, Davies-Bouldin Index)

2.5 Neuro-Symbolic and Knowledge-Driven Approaches

Introduction to Neuro-Symbolic XAI

Use of external knowledge (ontologies, UMLS, knowledge graphs)

Semantic annotation for enhancing explanation comprehensibility

Examples from literature using semantic knowledge for explanation

2.6 Explainability in Clinical Prediction Models

Survey of XAI applied to kidney transplant prediction and related medical domains

Challenges identified in clinical XAI:

Audience-centric needs (clinicians vs. developers)

Lack of evaluation benchmarks

Trustworthiness and causality

2.7 Gaps in Existing Literature

Lack of work combining process mining and surrogate models for XAI

Limited work using clustering to simplify explanation complexity

Need for domain-contextualized, semantically-rich explanation pipelines

Explainable AI (XAI): Summarize key XAI definitions (local/global, comprehensibility, fidelity, etc.).

Surrogate Modeling in XAI: Detail model-agnostic post-hoc explanation methods (LIME, SHAP, TREPAN).

Process Mining Concepts: Definitions of event logs, variants, trace clustering, and process discovery.

Clinical Decision Support in Transplantation: Short review of ML for kidney graft survival prediction.

Use: Jaber’s thesis chapters 1–2 and references for XAI theory.

## 2.1 Theoretical Background

## 2.2 Related Work

# Chapter 3 Methodology

3.1 Overview of the Research Approach

Introduce the overall neuro-symbolic XAI pipeline.

Distinguish between the two main components:

Data-driven component: extraction and optimization of decision paths from surrogate models.

Knowledge-driven component: semantic enrichment and structured explanation via process mining.

3.2 Data Sources and Case Study Context

Description of the SRTR kidney transplant dataset:

Features used (donor, recipient, procedural)

Outcome label (survival vs. failure)

Preprocessing steps (feature encoding, train-test split)

Justify why this clinical domain is well-suited for explainability research.

3.3 Surrogate Modeling of Black-Box ML Models

Train black-box models: Random Forest, XGBoost, and ANN.

Generate interpretable surrogate models: CART and TREPAN.

Objective: approximate predictions while enabling path extraction.

Key evaluation metrics: fidelity, accuracy, confidence.

3.4 Decision Path Extraction and Preprocessing

Extract decision paths from surrogate models.

Preprocess to improve interpretability:

Remove noisy or low-fidelity paths

Eliminate duplicates

Simplify and standardize logical conditions (e.g., transform negatives to positives)

Transform paths into event logs (each condition = event; order = synthetic timestamp)

3.5 Trace Clustering for Path Abstraction

Represent decision paths as traces for process mining.

Use TF-IDF embedding to convert traces into feature vectors.

Apply clustering (e.g., spectral clustering, k-means)

Rationale: reduce variability, increase interpretability

Choose optimal number of clusters using internal validation indices

3.6 Process Discovery and Model Construction

Apply process discovery (e.g., Heuristic Miner) on clustered traces.

Generate process models per cluster to reveal interpretable logic.

Highlight the trade-off between model complexity and path variability.

Visualize pathways leading to survival vs. non-survival outcomes.

3.7 Semantic Enrichment of Decision Paths

Use UMLS and other medical ontologies for semantic annotation.

Map features to domain-specific concepts to improve human interpretability.

Construct a Semantic Explanation Knowledge Graph (SeE-KG):

Nodes = concepts; edges = logical or clinical relations

Paths = conceptually enriched explanations

3.8 Explanation Generation and Interaction

Enable local and global explanations:

Local: trace back the path for a single prediction

Global: visualize key patterns across multiple predictions

Support interactive querying (e.g., explore alternative paths, ask “what-if” questions)

Discuss potential for human-in-the-loop review and expert validation

3.9 Summary of the Methodology

Recap the end-to-end pipeline from data preparation to explainable model output.

Emphasize novelty:

Integration of process mining into XAI

Combination of statistical, symbolic, and semantic techniques

3.1 Dataset & Black-box Models

Describe the kidney transplant dataset (SRTR), feature engineering, and ML models (RF, ANN, XGB).

3.2 Surrogate Model Extraction

Explain how CART trees were used to mimic black-box models.

Include your fidelity/confidence filtering and preprocessing steps.

3.3 Event Log Creation

How decision paths were transformed into process traces (with synthetic timestamps).

Discuss assumptions like treating feature order as logical sequence.

3.4 Trace Clustering

Explain use of TF-IDF, spectral clustering, and rationale for choosing cluster numbers.

#### 3.5 Process Discovery

Discuss use of Heuristic Miner and other miners (e.g., Fuzzy, Inductive).

Mention cases where spaghetti models appeared and how clustering helped.

Use: Methods section in your paperand slides on clustering and event log design.

# Chapter 4 Experiments and Evaluation

4.1 Experiment Design Overview

Describe the goal of the experiments:

Validate whether process mining and surrogate models can produce coherent and interpretable explanations of kidney transplant predictions.

Briefly introduce the experimental pipeline:

Model training → Surrogate modeling → Path extraction → Event log generation → Clustering → Process discovery → Evaluation.

4.2 Dataset and Experimental Setup

Dataset: SRTR kidney transplant dataset.

Number of cases (e.g., 6,306 survival, 7,363 failure).

Feature types (recipient, donor, procedural).

Preprocessing: feature selection, encoding, stratified train-test split.

Experimental Conditions:

Binary classification target: transplant survival vs. non-survival within 5 years.

Evaluation on held-out test data.

4.3 Black-Box and Surrogate Model Training

Black-box models: Random Forest, XGBoost, Artificial Neural Networks.

Training details: cross-validation, hyperparameter tuning.

Performance metrics: AUROC, Precision, Recall, F1-score.

Surrogate models: CART, TREPAN.

Goal: approximate predictions for interpretability.

Report fidelity (how well they mimic the black-box model).

4.4 Decision Path Extraction and Preprocessing

Number of paths extracted per model/class (e.g., 802 survival, 1336 non-survival).

Preprocessing steps:

Elimination of duplicates

Simplification and condition transformation

Creation of event log with synthetic timestamps

Result: a cleaned and structured event log suitable for process mining.

4.5 Trace Clustering and Analysis

TF-IDF encoding of traces for clustering.

Clustering algorithms used: spectral clustering (primary), k-means (comparison).

Evaluation criteria:

Internal validity metrics (Silhouette Score, Calinski-Harabasz, Davies-Bouldin).

Selected number of clusters (e.g., k = 10–12).

Outcome: clusters of decision paths with reduced variability and enhanced interpretability.

4.6 Process Discovery and Model Interpretation

Discovery algorithms used: Heuristic Miner (primary), Alpha/Fuzzy/Inductive Miners (possible alternatives).

Global model results:

Highly complex "spaghetti" diagrams (low interpretability).

Cluster-level models:

Simpler, structured, and visually comprehensible.

Show different decision strategies for survival vs. non-survival.

Visual examples: process models from high-precision clusters (e.g., Cluster 5 for survival).

4.7 Quantitative Evaluation of Interpretability

Use of process mining metrics:

Precision: how well the model restricts behavior.

Recall (Fitness): how well the model supports observed traces.

F1-score: harmonic mean of precision and recall.

Summary of results (from Table 1 in your paper):

Global models: low F1 (e.g., 0.08 survival, 0.09 non-survival).

Clustered models: much higher F1 (e.g., up to 0.96).

Comparison between clusters: interpretability vs. complexity.

4.8 Qualitative Evaluation and Insights

Clinical plausibility:

Clusters reflect meaningful patient subgroups (e.g., functional status, donor-recipient profiles).

Common vs. rare variants: identify typical vs. exceptional decision logic.

Interpretability impact:

Enables clinicians to view high-level strategies.

Reduces need to examine hundreds of decision trees.

Visual inspection of clusters reveals linear, decision-point-based logic.

4.9 Limitations of the Current Experiments

Synthetic timestamps (lack of real temporal data).

Some clusters still complex despite clustering.

Interpretability not yet evaluated with real clinical users (pending human studies).

Event coverage vs. accuracy trade-offs.

4.10 Summary of Experimental Findings

Demonstrated feasibility of using process mining to structure and simplify decision paths from multiple surrogates.

Clustering proved essential for extracting interpretable models.

The methodology supports both quantitative and qualitative gains in explainability.

# Chapter 5 Conclusion and Future Work

5.1 Summary of the Research

Recap the core goal: improving the interpretability of machine learning predictions in clinical settings (specifically kidney transplant outcomes) using process mining on surrogate decision paths.

Outline the main components of your methodology:

Surrogate modeling for black-box explanation

Decision path extraction and preprocessing

Event log creation and trace clustering

Process discovery to reveal structured decision logic

Emphasize novelty:

Reversing the usual XAI-process mining flow (using process mining for explainability, not the other way around)

Bridging black-box ML and human-understandable pathways

5.2 Key Findings

Technical outcomes:

Global process models were uninterpretable due to complexity; clustering made meaningful patterns visible.

Cluster-specific models showed higher fidelity and interpretability.

Trace clustering significantly improved model quality (F1-score improvements: e.g., from 0.08 → 0.52 on survival paths).

Clinical relevance:

Process models reflected medically plausible workflows.

Clinicians could inspect simplified clusters instead of hundreds of rules.

Rare vs. common paths allowed identification of potential biases or exceptional cases.

5.3 Contributions to the Field

Introduced a new hybrid framework combining XAI, surrogate modeling, and process mining.

Demonstrated the utility of trace clustering for organizing fragmented surrogate outputs.

Created a methodology applicable beyond kidney transplant prediction (e.g., in any domain with ensemble surrogates).

Advanced explainability through structured, aggregated, and semantically enriched outputs.

5.4 Limitations

Data-related:

No real temporal data—synthetic timestamps were used to order decision conditions.

Dataset may not fully represent rare clinical cases.

Methodological:

Clustering quality is dependent on parameter choices (e.g., TF-IDF, k value).

Surrogate models still imperfect approximations.

User validation:

Interpretability was evaluated qualitatively and through mining metrics, but no formal clinician user study was performed.

5.5 Directions for Future Work

Improving clustering and process discovery:

Explore alternative encodings beyond TF-IDF (e.g., n-gram embedding, graph embeddings).

Combine clustering with semantic similarity or medical domain heuristics.

Incorporate time and source metadata:

Use real timestamps if available or infer delays between conditions.

Add surrogate confidence, feature importance, or source model identifiers to event logs.

Expand semantic enrichment:

Automatically link decision conditions to clinical guidelines.

Add more layers to the Semantic Explanation Knowledge Graph (e.g., patient subtypes, outcomes).

Human-in-the-loop evaluation:

Conduct structured interviews or surveys with clinicians to assess interpretability, trust, and usefulness.

Let domain experts interactively adjust or refine the discovered models.

Generalization:

Apply framework to other clinical prediction tasks (e.g., ICU triage, cancer prognosis).

Test on other domains with ensemble ML (e.g., fraud detection, legal risk modeling).

5.6 Final Remarks

Emphasize that the work marks a shift from explaining individual predictions to explaining structured decision logic across many cases.

The framework balances fidelity and human interpretability, offering a new way to bridge the gap between ML systems and domain experts.

With further validation and refinement, it has the potential to become a key tool in transparent, trustworthy AI for healthcare.

## 5.1 Discussion

# References

[Use a reference manager to insert citations and bibliography]

# Appendix A