Narrative-based Explainable AI for Clinical Decision Making

University of Turku
Health Technology
Master of Science Thesis
Department of Computing
July 2025
Ziyun Pan

Supervisors:
Tapio Pahikkala
Daniele Raimondi (Institute of Molecular Genetics of Montpellier)
Francesco Codice (Institute of Molecular Genetics of Montpellier)

UNIVERSITY OF TURKU
Health Technology

Ziyun Pan: Narrative-based Explainable AI for Clinical Decision Making

Master of Science Thesis, 54 p.
Department of Computing
July 2025

The study presents a hybrid framework that integrates survival modeling, explainable AI, and large language models (LLMs) to generate interpretable narrative explanations for individual patients with multiple myeloma. A Random Survival Forest (RSF) is trained to predict median survival time using clinical features, with survival functions estimated accordingly. SHAP (SHapley Additive exPlanations) values are computed using SurvSHAP(t) at the predicted median survival time to quantify feature contributions. These SHAP values are then used to construct structured prompts that guide locally deployed LLMs in generating patient-specific explanations. To ensure privacy and offline capability, all LLMs are deployed locally using tools such as Ollama and vLLM. Experimental results show that the RSF provides superior predictive performance, and qualitative analysis of the LLM-generated outputs reveals variation across models in terms of factual alignment, reasoning quality, and linguistic fluency. Among all tested models, DeepSeek-R1 with 70B parameters produced the most coherent and clinically plausible explanations. This work demonstrates the potential of combining explainable survival models and LLMs for trustworthy, personalized AI interpretation in clinical settings.

Keywords: Survival Analysis, Explainable AI, SHAP, Random Survival Forest, Large Language Models, Narrative Explanations, Multiple Myeloma, Clinical Interpretation

Contents

List of Figures

List of Tables

List of acronyms

C-index Concordance Index

CoxPH Cox Proportional Hazards Model

DTD Deep Taylor Decomposition

iBS Integrated Brier Score

IPCW Inverse Probability of Censoring Weights

LIME Local Interpretable Model-agnostic Explanations

LLMs large language models

MM Multiple Myeloma

RF Random Forest

RSF Random Survival Forest

SHAP SHapley Additive exPlanations

VIMP Variable Importance

XAI explainable AI

1 Introduction

In recent years, machine learning (ML) models have been increasingly adopted in clinical settings to support diagnosis, prognosis, and treatment planning on tabular patient data (i.e. health records, clinical analysis, biomarkers). While traditional models like logistic regression and linear regression are still widely used due to their transparency, more complex models such as random forests and neural networks often outperform them in predictive accuracy. Specifically, survival models like random survival forest and cox model are used for survival data[1]. However, these advanced models are commonly perceived as "black boxes" because they offer limited interpretability—a critical shortcoming in high-stakes fields like healthcare.

To bridge this gap, explainable AI (XAI) techniques have been developed to shed light on model decision-making processes. Among them, SHAP (SHapley Additive exPlanations) has gained prominence for its strong theoretical foundation and ability to assign meaningful feature importance values for each individual prediction[2]. Nevertheless, SHAP's output—typically in the form of numeric tables or visual plots—can still be difficult for non-technical users such as patients, clinicians and healthcare stakeholders to fully comprehend and act upon.

Recent advancements in large language models (LLMs), such as GPT and DeepSeek, offer a new opportunity to enhance explainability. These models can generate fluent, natural language narratives that may be more accessible and intuitive than traditional plots or feature rankings. Narrative-based explanations are particularly effec-

tive in conveying scientific concepts to non-expert audiences, as storytelling has been shown to improve understanding, retention, and engagement in complex domains such as healthcare and science communication[3]. One research shows that integrating SHAP with LLMs enables the creation of patient-specific textual explanations that describe the reasoning behind each model prediction in plain language[4].

This project explores the combination of SHAP-based feature attribution with LLM-driven natural language generation to produce interpretable, instance-level narratives for clinical predictions. We specifically focus on patient outcome prediction in multiple myeloma, a complex hematologic malignancy. The proposed framework aims to improve transparency and trust in machine learning models by delivering explanations that are both technically accurate and human-readable. Ultimately, this approach seeks to empower clinicians and biomedical researchers with AI tools that provide not only reliable predictions but also clear, actionable insights.

1.1 Problem Definition

While machine learning models have demonstrated strong potential in clinical prediction tasks, their lack of interpretability presents a significant barrier for adoption among clinicians and patients. SHAP value has emerged as a powerful method for attributing model predictions to input features. However, SHAP value's outputs—typically expressed as numeric tables or complex plots—remain challenging for non-technical users to interpret. This limits the practical utility of these explanations in high-stakes healthcare contexts where clear, accessible communication is essential.

1.2 Research Questions

The project aims to improve the interpretability of machine learning predictions in clinical settings by integrating SHAP-based explanations with natural language narratives generated by large language models (LLMs). We address the following research questions:

1. Can survival models, such as Random Survival Forests, effectively predict clinical outcomes (e.g., relapse and overall survival) in multiple myeloma patients using structured clinical data, and can meaningful SHAP-based feature attributions be extracted to support interpretability at the individual level?

2. Can locally deployed large language models (LLMs) generate accurate, readable, and privacy-compliant narrative explanations of survival model predictions based on SHAP values, in accordance with regulations such as GDPR and the EU AI Act?

1.3 Thesis content summary

The rest of the thesis is organized as follows. Chapter 2 provides background knowledge and related work on survival analysis, explainable AI techniques, large language models (LLMs), and data security considerations. Chapter 3 presents a detailed overview of the multiple myeloma dataset, including feature selection and outcome variables. Chapter 4 describes the methodology, covering data preprocessing, model training, evaluation strategies, SHAP-based explanation, and the narrative generation pipeline. Chapter 5 presents the results and discussion, including model performance, interpretability, narrative analysis, and comparison of LLM responses. Chapter 6 concludes the thesis and outlines directions for future work. Finally chapter 7 illustrates the usage of AI in the thesis.

2 Background and Related Work

This chapter introduces the key concepts underlying this project, including survival analysis, model interpretation, narrative generation with LLMs and how we protect data security. These foundations support the goal of generating interpretable, personalized predictions for clinical outcomes in multiple myeloma patients.

2.1 Survival Analysis

Survival analysis refers to a set of statistical methods designed to analyze the time until the occurrence of a specific event, such as death, relapse, or disease progression. In cancer research and other clinical contexts, this event may represent various outcomes, including time from diagnosis to death or time from remission to relapse [1], [5].

A distinguishing feature of survival analysis is its ability to account for incomplete observations, where the event of interest has not occurred for all individuals by the end of the study period.

2.1.1 Time-to-event Data

In clinical studies, the data structure often involves tracking the time from a defined starting point (e.g., diagnosis or treatment initiation) to an event of interest. This is known as *time-to-event data*[6].

If the event occurs during the study, the exact time is observed and recorded. However, for individuals who do not experience the event within the study period or are lost to follow-up, the data are considered *censored*. The most common form is *right-censoring*, where the true event time is only known to exceed the observed duration.

2.1.2 Censoring Data

Censoring arises when the exact time of the event is unknown for some participants. This can occur for several reasons: the patient has not experienced the event by the end of follow-up, is lost to follow-up, or encounters another event that precludes further observation[1]. In such cases, only a partial survival time is observed, and the true time remains unknown. Standard statistical methods that ignore censoring can lead to biased results and inefficient estimates[5].

Recent research[7] demonstrates that increasing censoring rates adversely affect model performance, particularly for non-linear models such as the mixture density networks and random survival forests. The study shows that model performance deteriorates more significantly with higher censoring rates than with shorter observation windows, suggesting that censoring is a more critical factor in model robustness. These findings underscore the importance of evaluating models under varying censoring conditions to ensure their reliability in real-world clinical settings.

2.1.3 Survival Function

The survival function, often denoted as $S(t)$, represents the probability that a subject survives beyond a specified time t. Formally, it is defined as:

$$S(t) = P(T > t)$$

where T is the time to event (e.g., death or relapse). This function provides a comprehensive summary of time-to-event data by indicating the proportion of individuals expected to remain event-free at each time point. In the context of clinical studies such as those involving multiple myeloma, $S(t)$ gives insight into patient prognosis over time, helping to compare treatment groups or stratify patient risk levels.

Since not all individuals will experience the event during the study period, the survival function appropriately accommodates censored data, ensuring an unbiased estimate of survival probabilities[1]. As such, it plays a central role in both descriptive and inferential survival analysis.

2.1.4 Survival Models

Traditional machine learning models, such as random forests, support vector machines (SVM), and linear regression, are commonly used in various predictive tasks. However, these models are not directly applicable to survival analysis because the outcome variable in this context is time-to-event data, which often includes censored observations. Standard supervised learning frameworks do not natively handle censoring, making it necessary to adopt specialized survival models that can appropriately model both event times and censoring mechanisms.

Traditional Methods

Traditional methods such as the Cox Proportional Hazards (CoxPH) model have been widely applied in clinical research. However, as biomedical datasets increasingly become high-dimensional—often with more features than samples—the assumptions and limitations of classical models become apparent. Although the CoxPH model can be interpreted using SHAP values to visualize the influence of individual features on the predicted risk [8], its underlying proportional hazards

assumption may not hold in complex or heterogeneous clinical settings.

As the number of features grows, the likelihood of high multicollinearity among covariates also increases. Severe collinearity can lead to unstable estimates of the regression coefficients (β), inflated standard errors, and unreliable p-values, ultimately reducing the interpretability of the model. Moreover, with a large number of covariates, the Cox model becomes more prone to overfitting, especially when the sample size is relatively small.

Additionally, the Cox model assumes a linear relationship between the covariates and the log hazard, and that the hazard ratios remain constant over time. In high-dimensional data, verifying these assumptions for all features becomes increasingly difficult. Fundamentally, the CoxPH model is a linear model and thus struggles to capture complex nonlinear relationships or higher-order interactions between features—unless such terms are explicitly engineered, which further increases the dimensionality and complexity.

To address these limitations, various extensions have been proposed, including regularized Cox models, partial least squares regression, and nonparametric approaches such as random survival forests (RSF) [9]. Regularized Cox models, for instance, incorporate feature selection and shrinkage techniques to improve generalization. However, in biomedical datasets, the dimensionality often remains high due to the biological complexity of the domain, making it challenging to reduce the number of relevant features without losing critical information.

Random Survival Forest

In recent years, nonparametric and flexible machine learning approaches have gained increasing attention, with the Random Survival Forest (RSF) model standing out as a particularly effective method. RSF [10] is an extension of Breiman's Random Forests tailored specifically for right-censored survival data. Like standard random

forests, RSF is an ensemble learning method that aggregates predictions from multiple decision trees to enhance generalizability and robustness. However, instead of predicting discrete class labels or continuous values, RSF is designed to estimate time-to-event outcomes, commonly expressed as cumulative hazard functions (CHF) or survival functions $S(t)$.

Unlike regression-based survival models, RSF makes no distributional assumptions and can effectively capture nonlinear relationships and higher-order feature interactions. This flexibility makes RSF particularly suitable for exploratory survival analysis where prior knowledge about feature relevance may be limited. Additionally, RSF is known for its robustness to multicollinearity and overfitting, and performs well in high-dimensional settings. It also supports both categorical and continuous covariates and handles missing values natively [9].

Empirical studies have demonstrated the effectiveness of RSF over traditional survival models. For example, Senevirathne et al. [11] applied RSF to predict harvest timing in mixed-species forests, reporting significantly improved accuracy over CoxPH based on AUC and Brier Score metrics. Similarly, Germer et al. [8] compared four survival models—CoxPH, RSF, DeepSurv, and TabNet—on lung cancer prognosis. Using TNM staging features and MissForest imputation, RSF achieved the highest Concordance Index (0.703 ± 0.004) and the best Integrated Brier Score (0.145 ± 0.004), outperforming both traditional and deep learning methods. Importantly, RSF also retained interpretability via SHAP (SHapley Additive exPlanations), which revealed clinically meaningful risk factors.

The RSF algorithm operates as follows: First, B bootstrap samples are drawn from the original dataset. For each sample, a survival tree is grown by recursively splitting nodes based on a subset of features, selected at random, using a splitting rule that maximizes survival differences (e.g., the log-rank statistic). Trees are grown fully, with the constraint that each terminal node must contain at least $d_0 > 0$ unique

event instances, ensuring reliable survival estimation.

For each terminal node, the cumulative hazard function is estimated using the Nelson–Aalen estimator:

$$\hat{H}(t) = \sum_{t_i \leq t} \frac{d_i}{n_i}$$

where d_i is the number of events at time t_i, and n_i is the number of individuals at risk just prior to t_i. The ensemble CHF is then calculated by averaging the CHFs from all trees, and the survival function is obtained as:

$$\hat{S}(t) = \exp\left(-\hat{H}(t)\right)$$

Taken together, RSF offers a powerful and interpretable alternative to traditional survival analysis models. Its ability to deliver accurate predictions while retaining clinical transparency makes it highly suitable for biomedical applications. In this study, RSF was used to model survival outcomes in patients with multiple myeloma, benefiting from its ability to handle high-dimensional, right-censored clinical data and to provide interpretable survival estimates and feature importance rankings.

2.2 Explainable AI and SHAP

As machine learning models increase in complexity, their interpretability often diminishes. This presents a critical challenge in high-stakes domains such as healthcare, where understanding model predictions is essential for safety, accountability, and trust. Traditional models like linear regression and decision trees are inherently interpretable but may lack the predictive power of more complex models such as ensemble methods and deep neural networks [12]. This trade-off between accuracy and interpretability has sparked growing interest in developing methods that balance both.

Model transparency is not only a technical concern but also a regulatory and ethical one. For example, the European Union's General Data Protection Regulation (GDPR) establishes a "right to explanation" for automated decisions [13]. Similarly, initiatives such as DARPA's Explainable AI (XAI) program and national efforts in countries like China underscore a global push toward algorithmic accountability and transparency [14].

Explainable Artificial Intelligence (XAI) aims to address these demands by developing methods that make machine learning decisions more transparent, interpretable, and trustworthy. XAI enables users—both experts and non-experts—to understand the reasoning behind model outputs, facilitating ethical deployment and informed decision-making [15].

Several XAI techniques have been proposed, including LIME (Local Interpretable Model-agnostic Explanations), SHAP (SHapley Additive exPlanations), and Deep Taylor Decomposition (DTD) [15]. Among them, SHAP stands out for its solid theoretical grounding in game theory and its consistency across various model types. In this study, we adopt SHAP as the primary explanation method due to its ability to provide both global and local interpretability in a mathematically principled and model-agnostic way.

2.2.1 SHAP Value

SHAP (SHapley Additive exPlanations) is a model-agnostic framework for interpreting predictions. It assigns each feature an importance score based on the Shapley value, a concept from cooperative game theory. SHAP is the unique additive explanation method that satisfies local accuracy, missingness, and consistency[12].

SHAP belongs to a class of models that explain predictions as a sum of individual feature contributions. Formally, the model output is represented as:

$$f(x) = \phi_0 + \sum_{i=1}^{M} \phi_i \tag{2.1}$$

where:

- $f(x)$ is the model output for input x,
- ϕ_0 is the expected value of the model output when no features are known (i.e., the base value),
- ϕ_i is the SHAP value, representing the contribution of feature i,
- M is the total number of input features.

SHAP values are defined as the Shapley values of the conditional expectation function of the model:

$$f_x(z') = \mathbb{E}[f(z) \mid z_S], \tag{2.2}$$

where z' is a binary vector indicating the presence of each feature, S is the subset of present (non-zero) features in z', and z_S denotes the observed values of those present features.

Intuitively, SHAP estimates how much each individual feature contributes to the prediction by computing the average marginal contribution of that feature across all possible subsets of features. In other words, it measures how the model output changes when a given feature is included versus excluded from the input.

Despite its strong theoretical foundation, the exact computation of SHAP values is computationally expensive due to the exponential number of feature subsets involved. To make SHAP practical for real-world use, especially in complex models, several efficient approximation methods have been proposed—most notably, TreeSHAP for tree-based models.

2.2.2 TreeSHAP

While deep learning models excel in domains such as image classification, speech recognition, and natural language processing, tree-based models often outperform them on tabular data — a data type commonly encountered in clinical and biomedical settings. In many practical applications involving structured features, ensemble models such as gradient-boosted trees and random forests remain the most accurate and interpretable choices[2].

To provide model-agnostic explanations, SHAP offers a unified approach based on game theory[12]. However, the exact computation of SHAP values is exponential in the number of input features, making it computationally infeasible for high-dimensional models[2].

TreeExplainer addresses this limitation by enabling fast, exact SHAP value computation for tree-based models. It exploits the internal structure of decision trees to reduce the computational complexity from exponential to low-order polynomial time. This breakthrough bridges theory and practice by allowing SHAP values to be used efficiently in real-world tree ensembles[2]. TreeExplainer enables transparent, local explanations for predictions made by tree-based models. It decomposes the model output into additive contributions from each input feature, thereby turning black-box predictions into interpretable forms. Rather than treating the model as an opaque system, TreeExplainer quantifies how much each feature pushes the prediction higher or lower, while preserving key theoretical properties such as local accuracy and consistency.

Beyond feature-level explanations, TreeExplainer also supports SHAP interaction values, which generalize SHAP values to quantify feature interactions. These values are derived from the Shapley interaction index, allocating attribution not just to individual features, but also to pairs of features. The result is a matrix of attributions: diagonal elements represent main effects, while off-diagonal elements

capture pairwise interaction effects.

By enabling fine-grained, efficient, and theoretically grounded explanation of tree ensemble predictions, TreeExplainer serves as a powerful tool for model interpretability — particularly in high-stakes applications such as clinical decision support[2].

2.2.3 Local Explanation with SHAP

Although Random Survival Forests (RSFs) provide global variable importance measures such as Variable Importance (VIMP), which are useful for identifying features associated with the outcome in complex and high-dimensional data [10], [16], they do not offer instance-level interpretability. As a result, providing local explanations for individual predictions remains an open challenge.

To address this need, model-agnostic approaches such as LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) have gained prominence. In this study, we adopt SHAP due to its strong theoretical foundation, consistency, and practical suitability for interpreting tree-based models at the instance level[12]. In this project, local interpretability is particularly important, as we aim to understand which features most strongly influence the risk prediction for each individual patient. Given the variability in treatment effects and disease progression across individuals, personalized explanations are essential for clinical relevance.

SHAP was chosen to interpret the tree-based models employed in this study due to its ability to provide consistent, locally accurate feature attributions grounded in cooperative game theory [17]. Unlike earlier tree-specific methods (e.g., Saabas), which consider only a single ordering of features, SHAP values average over all possible feature permutations. This property ensures consistency: if a model changes such that a feature has a larger impact, its SHAP value does not decrease.

In addition to accuracy and theoretical soundness, SHAP offers powerful visualization tools. Summary plots illustrate global feature importance by showing the distribution of SHAP values across all instances, while dependence plots reveal how the SHAP value of a feature changes with its actual value, including interactions with other variables.

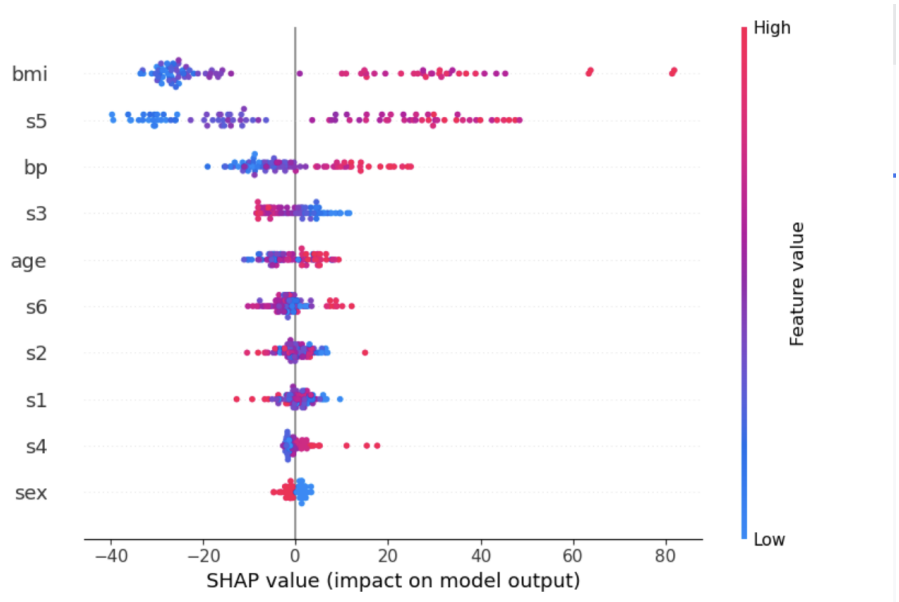

Figure 2.1: An example of feature importance visualization using SHAP values on diabetes dataset

2.3 Narrative Generation with LLMs

Large Language Models (LLMs) have recently demonstrated remarkable capabilities across a wide range of natural language processing (NLP) tasks, including but not limited to machine translation, text summarization, question answering, and dialogue generation. Built upon transformer architectures and trained on massive corpora, LLMs possess a sophisticated ability to understand contextual language patterns and to generate coherent, contextually appropriate responses [18]. Their emergence has reshaped the landscape of computational linguistics and AI-driven interaction, leading to a surge in academic and industrial interest.

One of the promising applications of LLMs lies in their capacity to bridge the gap between complex machine learning outputs and human interpretability. In the biomedical domain, where many predictive models rely on high-dimensional, domain-specific data, LLMs can be leveraged to translate model outputs into natural language narratives that are more accessible to end users, including clinicians, researchers, and even patients. This aligns with the broader goal of explainable artificial intelligence (XAI), which emphasizes not only transparency and trustworthiness, but also communication effectiveness.

In the context of this project, LLMs are used as a downstream component that interprets the output of a model trained on survival data. After model predictions and SHAP value analyses are obtained, the relevant clinical features and their contributions are embedded into natural language prompts. These prompts are then used to generate patient-specific narrative explanations that reflect both the statistical reasoning of the model and the clinical context of the data. This approach enhances the accessibility of predictive modeling, supports informed decision-making, and contributes to the reliability and adoption of AI systems in healthcare.

2.3.1 LLMs for Explaining AI

Recent studies, such as [4], suggest that narrative-driven explanations generated by LLMs (e.g., ChatGPT and DeepSeek) can significantly enhance the interpretability of complex model predictions. For instance, instead of relying solely on abstract visualizations like SHAP value plots, the transformation of feature attributions into structured, human-readable stories has shown to be more persuasive and intuitive for non-expert audiences. These narrative explanations, referred to as "SHAP stories," help users grasp not only which features are important, but also how these features interact in determining the outcome.

Local explanation techniques are essential for elucidating why a specific output

(e.g., token, label) is produced given a particular input. Unlike global interpretability, local methods aim to provide instance-specific reasoning. These techniques fall into several categories: feature attribution (e.g., SHAP, Integrated Gradients), attention-based explanation, example-based reasoning (e.g., counterfactuals), and natural language generation. Among these, natural language explanations generated by LLMs offer a user-friendly and accessible way to communicate the rationale behind model decisions, particularly in high-stakes domains such as clinical decision support. When used in combination with quantitative methods like SHAP, LLMs can effectively translate complex feature attributions into coherent narratives tailored to individual cases, bridging the gap between model transparency and human interpretability[19].

2.3.2 Effectiveness of SHAPstories

A research[4] evaluated SHAPstories—narrative-style explanations derived from SHAP values—for their effectiveness in enhancing interpretability of AI predictions. Across three user surveys, involving both general audiences and data science experts, SHAPstories consistently outperformed traditional SHAP plots in terms of perceived convincingness, user experience, and comprehension accuracy.

In qualitative assessments, 93.2% of general users and 77.8% of data scientists found SHAPstories to be more convincing than SHAP plots in explaining model outputs. Furthermore, 81.4% of experts considered SHAPstories a valuable supplement to SHAP visualizations. In terms of usability, general users reported higher ease of interpretation (92.4%), increased confidence (79.7%), and greater likelihood of future use (92.2%). Data scientists showed moderate enthusiasm for personal use but rated SHAPstories highly when intended for non-expert communication (91.7% agreed they would aid understanding for general users).

A quantitative evaluation in a credit scoring context further demonstrated that

SHAPstories significantly improved comprehension. For example, participants achieved 83.8% accuracy in summarizing a model's decision using SHAPstories, compared to only 33.8% with SHAP plots. This trend held across various tasks, with statistically significant improvements observed in multiple cases ($p < 0.05$).

Although SHAPstories required slightly more time to read (on average 10.6% longer), this additional effort correlated positively with improved decision accuracy. In contrast, time spent with SHAP plots did not enhance comprehension and even reduced confidence among users.

Qualitative feedback from experts suggests that SHAPstories are particularly useful for translating technical model outputs into human-friendly explanations. They help bridge the gap between visual summaries and narrative understanding, especially for non-technical stakeholders. While some experts noted that stories may occasionally overextend or oversimplify, the consensus recognized their value as a communicative aid rather than a replacement for SHAP plots[4].

This provides strong motivation for adopting SHAPstories, generated by LLMs, as the primary explanation method in this work.

2.4 Data Security and Privacy

The use of machine learning and large language models (LLMs) in clinical settings necessitates rigorous attention to data security and privacy. Clinical datasets often contain sensitive patient health information, and as such, their use is governed by strict regulatory frameworks. Among the most relevant are the General Data Protection Regulation (GDPR) in the European Union, the Health Insurance Portability and Accountability Act (HIPAA) in the United States, and the recently introduced European Union Artificial Intelligence Act (EU AI Act)[13], [20], [21].

In this project, all patient data is anonymized and handled exclusively in a secure, local computing environment. No cloud-based services or third-party APIs are used

at any stage of data processing or model inference. The LLM used for generating natural language explanations is also locally deployed, thereby eliminating the risk of data leakage associated with external API calls. Furthermore, at no point during the workflow is any data uploaded to public LLM platforms, including web-based or client-hosted interfaces. All interactions with language models are conducted entirely within a secure, offline environment.

The EU AI Act classifies AI systems used in healthcare as "high-risk," requiring such systems to demonstrate transparency, accountability, robustness, and appropriate human oversight. By ensuring that all components of the pipeline—data storage, model training, and LLM-based explanation—are locally contained and auditable, the project aligns with these regulatory expectations. This setup not only safeguards patient privacy but also reinforces the reliability and ethical soundness of AI-assisted clinical decision support.

3 Dataset

The clinical dataset used in this project originates from the research group of Dr. J. Moreaux (IGH/CHU Montpellier) and focuses on patients diagnosed with multiple myeloma (MM) and related hematological conditions between 2006 and 2022. The dataset contains detailed clinical, biological, and treatment-related information. Due to the sensitivity of patient data, the dataset is not publicly available. All experiments and analyses were conducted locally, and only locally deployed large language models (LLMs) were used to ensure compliance with data protection regulations.

3.1 Dataset Overview

Table 3.1 summarizes the differences between the original and filtered datasets used in this study.

	Original Dataset	**Filtered Dataset**
Number of Samples	239	157
Number of Features	44	29
Outcome Variables	EFS, EFSA, OSA, OS	EFSA, OSA
Numerical Features	✓	✓
Textual Features	✓	✗

Table 3.1: Overview of the original and filtered datasets used in the project.

Feature Overview It consists of a wide range of features, which can be grouped into the following categories:

Feature Category	**Description**
Patient identifiers and demographics	Patient ID, age, date of clinical entry.
Diagnostic information	Pathology classification (e.g., MM, MGUS, PCL), disease staging systems such as Salmon-Durie and ISS/R-ISS.
Genetic and molecular markers	IgH translocations, IgL types, and availability of RNASeq, RRBS, WES, WGS data.
Treatment history	Induction therapy received, number and type of stem cell transplants, treatment response outcomes.
Biomarkers and lab measurements	CD138+ cell percentage, β_2-microglobulin, albumin, LDH, CRP, calcium, immunoglobulin levels (IgG, IgA, IgM), hemoglobin, creatinine, proteinuria, osteolytic lesions, S-phase percentage.
Outcome variables	Event status and time-to-event data for relapse (EFS/EFSA) and overall survival (OS/OSA), including censoring indicators.

Table 3.2: Overview of feature categories included in the dataset.

3.2 Feature Selection

Feature selection was performed based on domain knowledge and clinical guidance provided by Dr. J.Moreaux. The inclusion criteria were as follows:

1. The dataset was filtered to retain only patients who underwent HDT+Auto

treatment (i.e., high-dose melphalan followed by autograft).

2. Biomarker-related variables were prioritized, as they provide objective measurements for disease burden and patient status.

3. Clinical staging and genetic features such as Stade_SD, Stade_ISS, IgH, and IgL were included due to their strong prognostic value in multiple myeloma.

After selecting features for further modeling, we summarized the filtered datasets for each outcome variable in Table 3.3.

Metric	**OSA** (Death without Allograft)	**EFSA** (Relapse without Allograft)
Number of Samples	154	152
Number of Events (1)	59	118
Number of Censored (0)	95	34
Censoring Rate	61.69%	22.37%

Table 3.3: Summary of filtered outcome variables: OSA and EFSA.

3.3 Additional Description

3.3.1 Features

The clinical feature names and their corresponding descriptions are presented below to facilitate interpretability and support subsequent use in prompting large language models (LLMs). These definitions also aid in understanding the clinical context of the data.

- **Age**: The patient's age at diagnosis. Age is a key prognostic factor that often influences treatment decisions and survival outcomes.

- **Stade_SD**: The Salmon and Durie staging system classifies tumor burden based on clinical symptoms, serum calcium levels, hemoglobin, presence of bone lesions, and monoclonal protein production. Stage I indicates low tumor burden, while Stage III indicates a high burden of disease[22].

- **Stade_ISS**: The International Staging System (ISS) uses serum 2-microglobulin and albumin levels to stratify patients into three stages: I (best prognosis), II (intermediate), and III (poor prognosis)[23].

- **poucentage_CD138_Moelle / %CD138 au myélogramme**: Percentage of CD138-positive cells in bone marrow aspirate. CD138 is a plasma cell surface marker used to identify malignant cells.

- **NPC_par_mm3**: Absolute number of nucleated plasma cells per cubic millimeter of bone marrow.

- **MMC_par_mm3 (Moelle)**: Malignant myeloma cells per mm^3 in the bone marrow, reflecting disease infiltration.

- **Pourcentage_Phase_S_PC_tumoraux**: The percentage of tumor plasma cells in the S-phase of the cell cycle, indicating proliferative activity.

- **Ostéolyse**: Presence of osteolytic bone lesions, a common manifestation of myeloma-related bone disease.

- **Tx_Sg_Prot_MonoCl**: Serum monoclonal protein level (M-protein), an important marker for disease burden and response to therapy.

- **Taux IgG / IgA / IgM g/L**: Levels of immunoglobulin G, A, and M in serum. Abnormal levels are associated with different subtypes of multiple myeloma.

- **Calcémie mmol/L**: Serum calcium concentration. Hypercalcemia is a common metabolic complication of advanced myeloma.
- **protéinurie g/24h**: 24-hour urinary protein excretion. Elevated values may indicate kidney damage or involvement.
- **Hémoglobine g/dL**: Hemoglobin level in blood. Anemia is a frequent feature of multiple myeloma due to bone marrow infiltration.
- **Béta-2M mg/L**: Beta-2 microglobulin level, a key component of the ISS and a marker of tumor load and renal function.
- **CRP mg/L**: C-reactive protein, a non-specific inflammatory marker that may correlate with disease activity.
- **LDH IU/L**: Lactate dehydrogenase, a metabolic enzyme that serves as a surrogate marker for tumor burden or cellular turnover.
- **Albumine g/L**: Serum albumin concentration, used in ISS and indicative of nutritional and systemic status.
- **Créatinine µmol/L**: Serum creatinine, reflecting renal function, which is often impaired in myeloma patients.
- **Nb greffes**: Number of stem cell transplants (autografts) received by the patient.
- **BJ**: Presence of Bence-Jones protein in the urine, indicative of light-chain multiple myeloma.
- **IgA / IgD / IgG / NS**: Immunoglobulin isotype classification of the monoclonal component. NS refers to "not specified."
- **Kappa / Lambda**: Type of light chain produced by malignant plasma cells. These are essential for identifying the clonal nature of the disease.

3.3.2 Outcome Variables

The dataset includes two main types of outcome variables:

- **EFSA** and **Days-EFSA**: These represent the event and the number of days until the first relapse, respectively. Both variables are already censored for allogeneic transplant (i.e., patients who received an allograft are treated as censored).
- **OSA** and **Days-OSA**: These indicate the event and the number of days until death. These variables are also censored for allogeneic transplant.

These outcomes are treated as survival targets in this project. EFSA (Event-Free Survival censored for allograft) is used to model the time to first relapse, and OSA (Overall Survival censored for allograft) for time to death.

4 Method

This project follows a structured pipeline to generate interpretable predictions for survival analysis. First, clinical data is collected and preprocessed to ensure quality and consistency. A Random Survival Forest (RSF) model is then trained to predict patient-specific survival functions. Next, SHAP values are computed to quantify the contribution of each feature to individual predictions. Finally, a locally deployed Large Language Model (LLM) generates narrative explanations (SHAPstories) based on the SHAP attributions, enabling user-friendly, interpretable output. An overview of the full workflow—is illustrated in Figure 4.1.

To ensure reproducibility and address potential compatibility issues, the key Python packages and their corresponding versions used in this project are summarized in Table 4.1. Due to the high computational demands of locally deployed large language models (LLMs), initial development and testing were conducted on a MacBook with an M3 Pro chip. Subsequently, LLM-related experiments were executed on a high-performance computing cluster. To maintain consistency across environments and ensure stable execution, the same package versions were installed on both platforms.

Table 4.1: Versions of Key Python Packages Used

Package	**Version**
numpy	1.24.3
survshap	0.4.2
shap	0.42.1
scikit-survival	0.22.2
scikit-learn	1.3.0

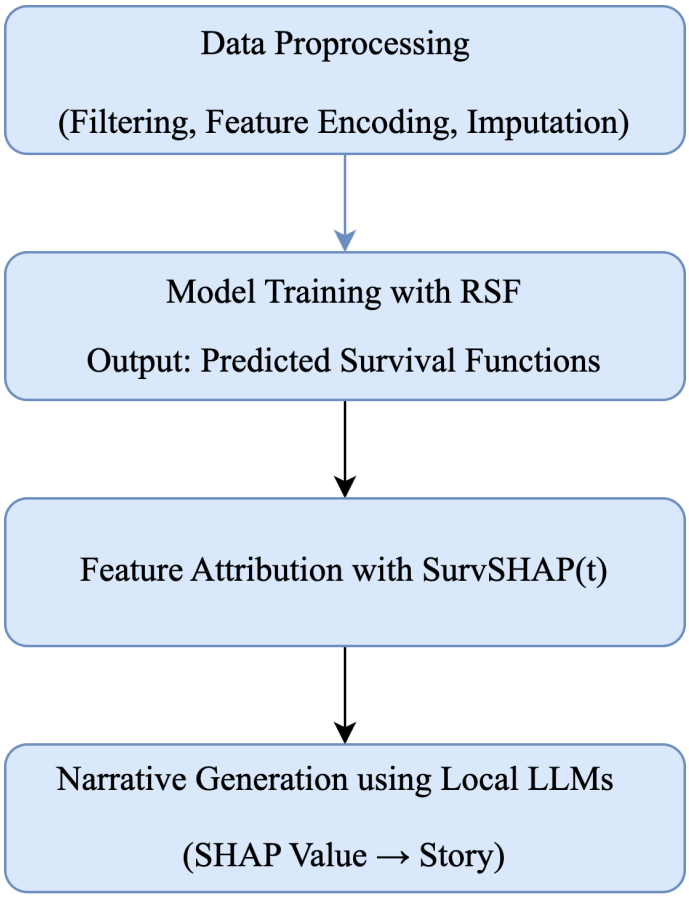

Figure 4.1: Overview of the workflow from data preprocessing to explainable prediction using SHAP and LLM-generated SHAPstories.

4.1 Data Preprocessing

We applied a comprehensive data preprocessing pipeline tailored to clinical data with mixed numeric, categorical, and textual formats. The preprocessing included the following key steps:

1. **Patient Filtering:** We retained only patients who had undergone autologous stem cell transplantation (`HDT+AUTO = Oui`) to ensure cohort consistency.

2. **Data Cleaning and Transformation:** Numerical fields with irregular string values (e.g., ''et'' denoting ranges or ''<'' indicating lower bounds) were cleaned and parsed into float format. Specific columns like `Taux IgA g/L`, `Taux IgM g/L`, and `Tx_Sg Prot MonoCl` were processed using custom parsing logic.

3. **Staging Variable Encoding:** Clinical staging variables, `Stade_SD` (Durie-Salmon system) and `Stade_ISS` (International Staging System), were mapped to ordinal numeric values according to standard medical definitions. An optional one-hot encoding alternative was also considered.

4. **Handling Categorical Variables:** Immunoglobulin gene alterations (`IgH`, `IgL`) were split and transformed using multi-label binarization (one-hot encoding). The binary variable `Ostéolyse` was mapped to 0/1.

5. **Missing Data Handling:** Columns with missing values exceeding a configurable threshold (default 30%) were dropped from the dataset. For the remaining numeric columns (excluding survival target and censoring indicator), missing values were imputed using an out-of-scale constant value of `-999999`. This imputation strategy leverages the robustness of Random Forests, which can effectively handle such extreme values without introducing bias, as the algorithm performs implicit data partitioning and is relatively insensitive to monotonic transformations or outliers.

6. **Target Definition and Final Filtering:** Two survival targets were available: `Days_EFSA` and `Days_OSA`, each with a corresponding censoring indicator. Based on the selected label (`osa` or `efsa`), we constructed the label and event vectors, then removed rows with missing survival or censoring information.

The result was a clean dataset suitable for survival modeling, with well-defined covariates (`X`), survival times (`y`), and event indicators (`y_censor`).

4.2 Model Training

To build a predictive model for survival outcomes, we used the Random Survival Forest (RSF) algorithm, implemented via `scikit-survival`. RSF is a non-parametric ensemble learning method that extends the random forest approach to right-censored survival data. It is robust to high-dimensional input and does not require the proportional hazards assumption.

The data was randomly split into training and test sets using an 80/20 ratio. Stratified sampling was applied based on the censoring indicator to ensure class balance. The survival outcome was encoded as a structured array using `Surv.from_arrays()`, combining both time-to-event and censoring status(event indicator) into a format compatible with `scikit-survival`'s estimators.

A grid search was conducted to identify the optimal hyperparameters for the RSF model. The result is shown in the table:

Table 4.2: Best hyperparameter settings for RSF models determined via grid search.

Hyperparameter	**Event of Death**	**Event of Relapse**
`n_estimators`	200	175
`min_samples_split`	2	2
`min_samples_leaf`	1	1
`max_features`	log2	log2

4.3 Evaluation

To evaluate the predictive performance of the Random Survival Forest (RSF) model on survival outcomes, we employed two widely used evaluation metrics in survival analysis: the Integrated Brier Score (IBS) and the Concordance Index (C-index). These metrics respectively capture the model's calibration and discrimination ability.

4.3.1 Brier Score

To assess the calibration and overall accuracy of the survival model over time, we adopt the time-dependent Brier Score (BS), a proper scoring rule that measures the squared difference between predicted survival probabilities and the observed outcomes at a specific time t [24]. In the presence of right-censored data, the Brier Score incorporates inverse probability of censoring weights (IPCW) using the Kaplan–Meier estimate of the censoring distribution[25]:

$$\mathrm{BS}^c(t) = \frac{1}{n}\sum_{i=1}^{n}\left[\mathbb{I}(y_i \le t \wedge \delta_i = 1)\cdot\frac{(0-\hat{S}(t,\mathbf{x}_i))^2}{\hat{G}(y_i)} + \mathbb{I}(y_i > t)\cdot\frac{(1-\hat{S}(t,\mathbf{x}_i))^2}{\hat{G}(t)}\right] \tag{4.1}$$

where:

- $\hat{S}(t, \mathbf{x}_i)$ is the predicted survival probability for subject i at time t,
- y_i is the observed time (event or censoring),
- δ_i is the event indicator: 1 if event occurred, 0 if censored,
- $\hat{G}(t)$ is the Kaplan–Meier estimate of the censoring survival function,
- $\mathbb{I}(\cdot)$ is the indicator function.

To address right-censoring in survival analysis, we use the Inverse Probability of Censoring Weighting (IPCW) when computing the Brier Score. The standard Brier Score measures the squared difference between the predicted survival probability $\hat{S}(t, \mathbf{x}_i)$ and the actual outcome at time t. However, in survival data, some events are unobserved due to censoring (e.g., patients lost to follow-up), which can bias this score.

IPCW corrects for this by reweighting each individual's contribution based on the inverse of their probability of remaining uncensored until time t, denoted as $\hat{G}(t)$. This censoring distribution is typically estimated using the Kaplan–Meier estimator, a non-parametric method for estimating survival probabilities from censored data. The idea is that uncensored individuals provide more reliable information and should therefore be weighted more heavily.

Specifically, for patients who experienced the event before or at time t, the squared error is divided by $\hat{G}(y_i)$; for those still at risk after t, it's divided by $\hat{G}(t)$. This ensures fair evaluation of model performance despite uneven or informative censoring.

To assess performance across the entire follow-up time, we compute the Integrated Brier Score (IBS) as the weighted average of Brier scores over a time interval $[t_1, t_m]$, using a weight function $w(t) = \frac{t}{t_m}$ to emphasize later time points:

$$\text{IBS} = \int_{t_1}^{t_m} \text{BS}^c(t) \cdot w(t)\, dt \tag{4.2}$$

The IBS is conceptually similar to the integrated time-dependent AUC (iAUC), as both aggregate model performance across time by integrating time-specific metrics. Whereas iAUC captures discrimination, IBS reflects both calibration and discrimination. Mathematically, IBS represents the area under the Brier score curve across the follow-up period, normalized by the length of the time interval[26].

Lower IBS values indicate better calibration and predictive accuracy across the full time horizon. In this study, we report the mean IBS across 10-fold cross-validation to ensure robust evaluation.

4.3.2 C-index(ipcw)

To evaluate the discriminative performance of the survival model, we use the time-dependent Concordance Index (C-index), adjusted for censoring through Inverse Probability of Censoring Weighting (IPCW). The C-index measures the model's ability to correctly rank survival times. For a pair of individuals, the model is concordant if the patient with a shorter observed survival time is also predicted to have a higher risk (i.e., lower survival probability).

The IPCW-adjusted C-index is defined as:

$$\widehat{C}_{\text{IPCW}} = \frac{\sum_{i,j} \hat{W}_{i,j} \cdot \mathbb{I}(\hat{f}(\mathbf{x}_i) > \hat{f}(\mathbf{x}_j))}{\sum_{i,j} \hat{W}_{i,j}}, \tag{4.3}$$

where:

- $\hat{f}(\mathbf{x}_i)$ denotes the predicted risk score (e.g., negative survival probability).
- $\hat{W}_{i,j}$ is the inverse probability of censoring weight for the pair (i, j), estimated using the Kaplan–Meier estimator $\hat{G}(t)$ of the censoring distribution.
- $\mathbb{I}(\cdot)$ is the indicator function, which returns 1 if the predicted ordering of risk agrees with the observed outcome, and 0 otherwise.

In this study, we use the implementation provided in the `sksurv.metrics.concordance_index_ipcw` function from the scikit-survival library to compute the IPCW-adjusted C-index. We report the mean C-index across 10-fold cross-validation to ensure reliability and reduce variance.

4.4 Cross Validation

Two different cross-validation strategies were employed in this study. First, to determine the optimal threshold for missing data removal, we performed 10-fold cross-validation and evaluated model performance using the mean Integrated Brier Score (iBS) and Concordance Index (C-index) under each threshold setting. Second, to obtain robust patient-level survival function estimates, we used RepeatedStratifiedKFold to average the predicted survival functions across multiple resampled splits.

4.4.1 Missing Threshold Selection

Table 4.3: 10-fold Cross-Validation Results of RSF with Different Missing Value Thresholds

Target	Missing Threshold	Mean iBS	Mean C-index
osa	0.3	0.0781 ± 0.0305	0.6849 ± 0.1086
osa	**0.5**	**0.0776 ± 0.0324**	**0.7055 ± 0.1268**
osa	1	0.0792 ± 0.0326	0.6756 ± 0.1092
efsa	0.3	0.1378 ± 0.0226	0.6259 ± 0.1242
efsa	0.5	0.1369 ± 0.0249	0.6217 ± 0.1462
efsa	**1**	**0.1337 ± 0.0223**	**0.6514 ± 0.1486**

Based on the cross-validation results presented in Table 4.3, we selected different missing value thresholds for the two survival targets. For `osa`, a missing threshold of 0.5 yielded the best predictive performance, with the lowest mean iBS (0.0776 ± 0.0324) and highest mean C-index (0.7055 ± 0.1268). Therefore, columns with more than 50% missing values were removed for the `osa` target.

In contrast, for the `efsa` target, retaining all features (threshold = 1.0) resulted in superior model performance (mean iBS = 0.1337 ± 0.0223, C-index = 0.6514 ± 0.1486). As such, no feature elimination based on missingness was applied for the `efsa` prediction task.

4.4.2 Robust Survival Function Estimation

To obtain robust and reliable estimates of individual survival functions, we employed a repeated cross-validation strategy using the RepeatedStratifiedKFold method. In each fold, a survival model—such as Random Survival Forest (RSF)—was trained on the training subset and then used to predict the survival functions for the test instances.

For each patient, multiple survival probability curves were generated from the folds where the patient was included in the test set. These predictions were interpolated on a shared time grid and averaged, yielding an ensemble-based survival estimate that smooths out variability caused by different training-test splits.

To further enhance interpretability, we applied SurvSHAP(t), a time-dependent SHAP framework designed specifically for survival models. For each test instance, SurvSHAP(t) computes feature attributions across the entire survival timeline, thereby capturing how each variable contributes to the predicted survival probability at different time points. This allows for instance-level, temporally resolved explanations, which are essential in clinical contexts where feature influence may evolve over time.

This cross-validated setup ensures that both predictions and SHAP explanations remain unbiased, stable, and generalizable. The full pipeline is illustrated in Figure 4.2.

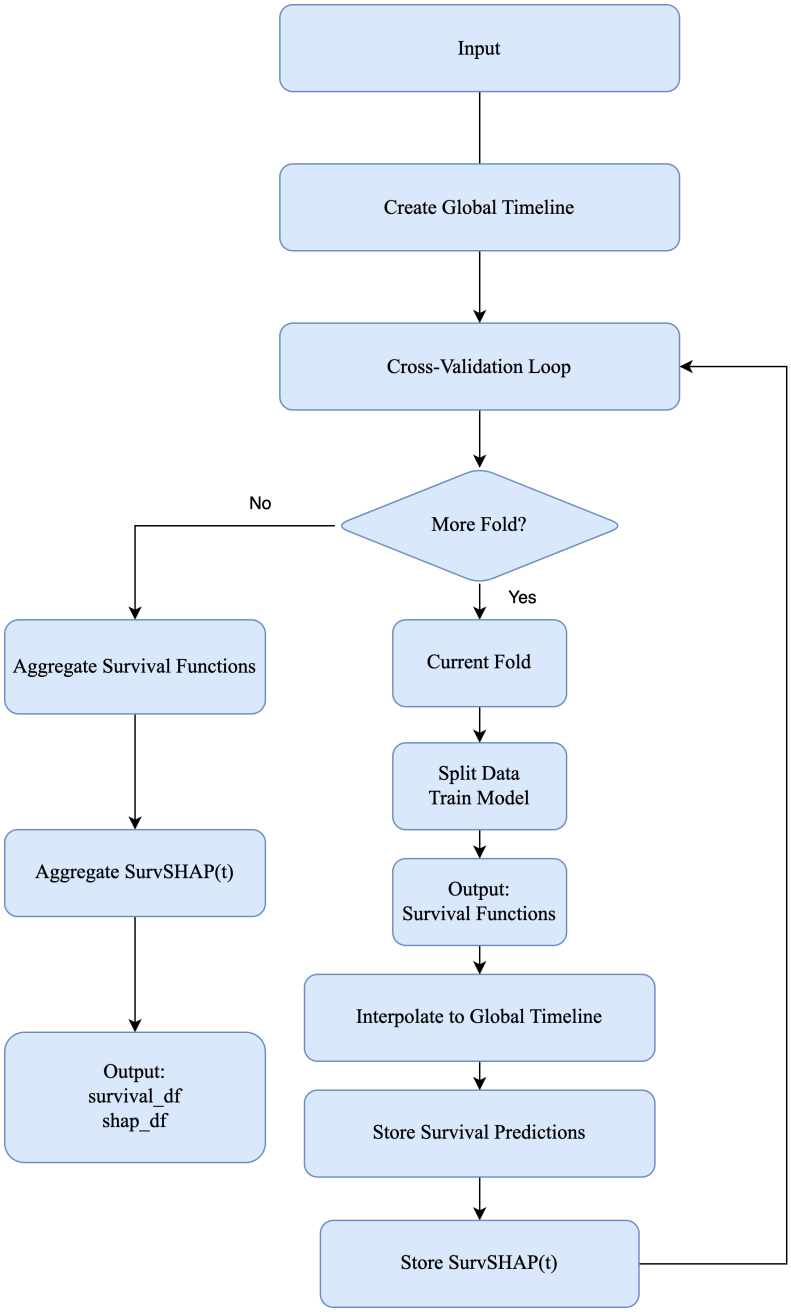

Figure 4.2: Overview of the cross-validation pipeline for aggregated survival functions and SurvSHAP(t).

4.5 Explanation with SHAP

4.5.1 SurvSHAP(t)

SurvSHAP(t) is a time-aware extension of the SHAP framework tailored for survival models. It generates individualized, time-dependent explanations by aligning SHAP values with the survival probability function over time. This method preserves the local accuracy property, ensuring that the cumulative contribution of features aligns with the model's predicted survival function at each time point.

Unlike static feature attribution methods, SurvSHAP(t) captures how the influence of each variable evolves temporally, allowing for a more nuanced understanding of feature effects across the survival horizon. Empirical studies have demonstrated that SurvSHAP(t) effectively identifies covariates with time-varying importance and provides more informative feature importance summaries compared to alternative approaches such as SurvLIME [27].

In this project, SurvSHAP(t) was used to generate patient-specific explanation curves, which help visualize not only which features contribute most to predicted risk, but also when these contributions become more pronounced over time.

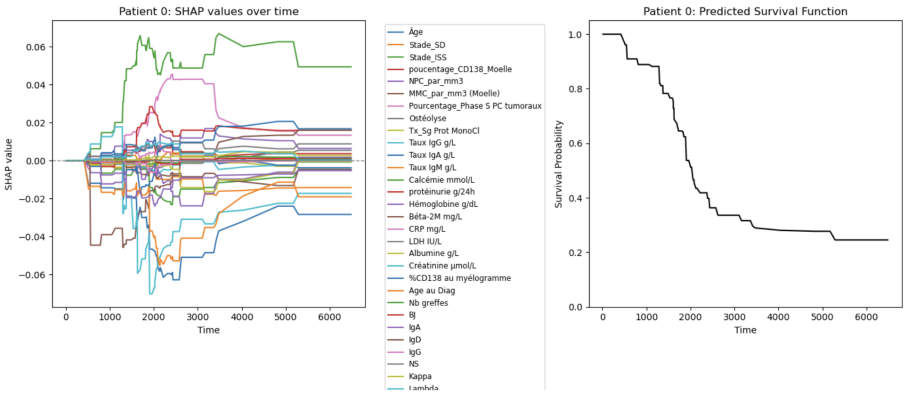

Figure 4.3: SurvSHAP(t) and survival function for Patient 1

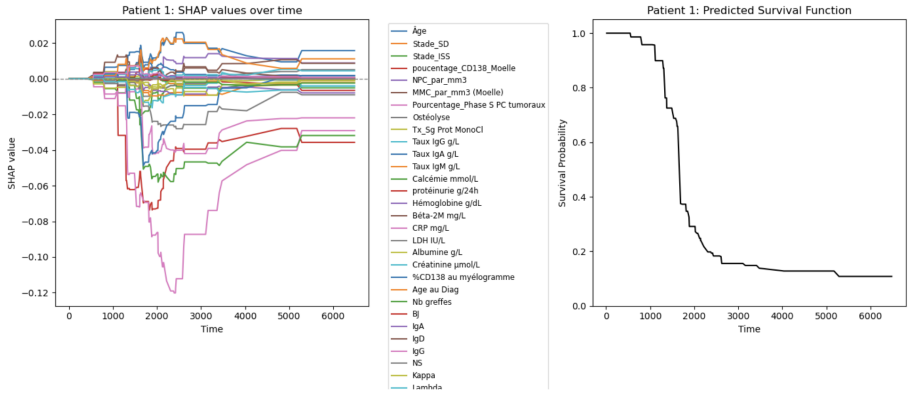

Figure 4.4: SurvSHAP(t) and survival function for Patient 2

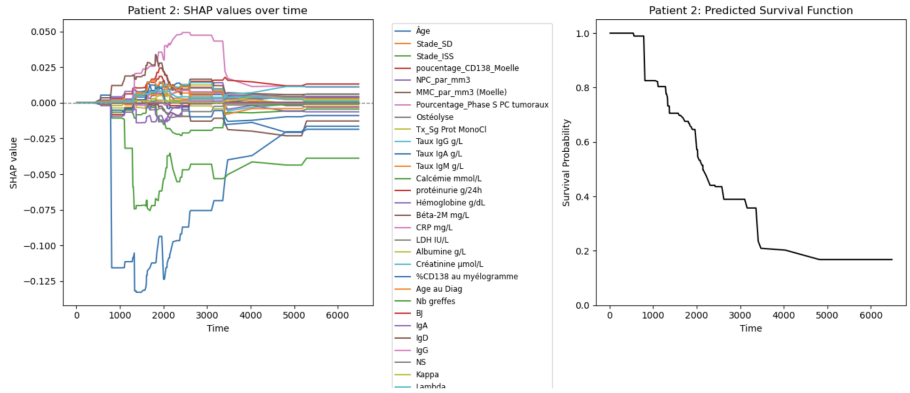

Figure 4.5: SurvSHAP(t) and survival function for Patient 3

4.6 Median Survival Time

In this study, median survival time is adopted as a key summary statistic to evaluate model-predicted survival functions. Compared to the arithmetic mean, the median offers several advantages in the context of right-censored survival data, which are commonly encountered in clinical research.

First, the median is less sensitive to extreme values and skewed distributions, making it a more robust estimator of central tendency in survival analysis. This is particularly relevant for patient populations exhibiting heterogeneous prognoses, where long right tails are often observed in survival times[28], [29].

Second, the interpretability of the median survival time enhances its clinical utility. It represents the time by which 50% of the population is expected to have experienced the event of interest (e.g., relapse or death). This interpretation aligns with common practices in clinical prognosis reporting and facilitates communication of results to both healthcare professionals and non-expert stakeholders[30].

Finally, the median can often be estimated reliably despite censoring, whereas the mean may not be computable when follow-up is incomplete. This robustness to censored observations makes the median particularly suitable as a comparative metric for survival models.

Accordingly, the median survival time is used in this project to quantify and compare survival outcomes at the individual level, based on the survival probabilities predicted by the model.

The process for extracting instance-level SHAP values at the median survival time is illustrated in Figure 4.6.

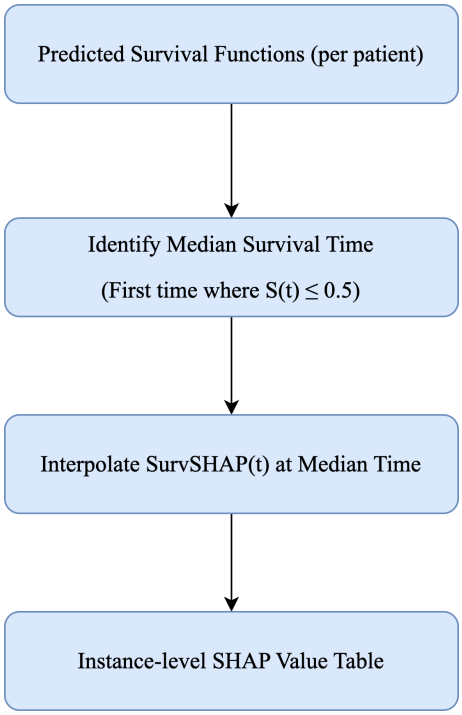

Figure 4.6: Pipeline for extracting instance-level SHAP values at the median survival time using SurvSHAP(t).

4.7 Narrative Interpretation

4.7.1 Narrative Generation Pipeline

To generate patient-level narrative explanations from model predictions, we adapted and extended the SHAPstory pipeline proposed by the research [4]. This approach leverages SHAP values to guide the construction of free-text rationales that interpret individual survival predictions. While the original SHAPstory framework was designed for classification tasks, we customized and expanded it to accommodate survival analysis.

We build on the previously extracted instance-level SHAP values, interpolated at each patient's predicted median survival time. To ensure interpretability, only features with valid values and absolute SHAP scores above a predefined threshold (e.g., 0.01) are retained. These filtered SHAP explanations serve as the foundation for generating narrative rationales.

Each retained feature is associated with a clinically grounded description derived from domain knowledge (e.g., "Hemoglobin (g/dL): Normal: 13–17 (male), 12–15 (female). Lower levels indicate anemia, a common complication."). These descriptions, along with each feature's SHAP score and observed value for the patient, are compiled into a prompt. The prompt provides context about the dataset, the predicted survival time, the actual clinical outcome (censored or not), and the SHAP-based rationale.

The process of generating structured prompts is illustrated in Figure 4.7. Based on instance-level SHAP values, the pipeline filters relevant features, matches them with predefined clinical descriptions, and composes a context-rich prompt. This prompt is then provided as input to the locally deployed large language models (LLMs) to produce narrative explanations.

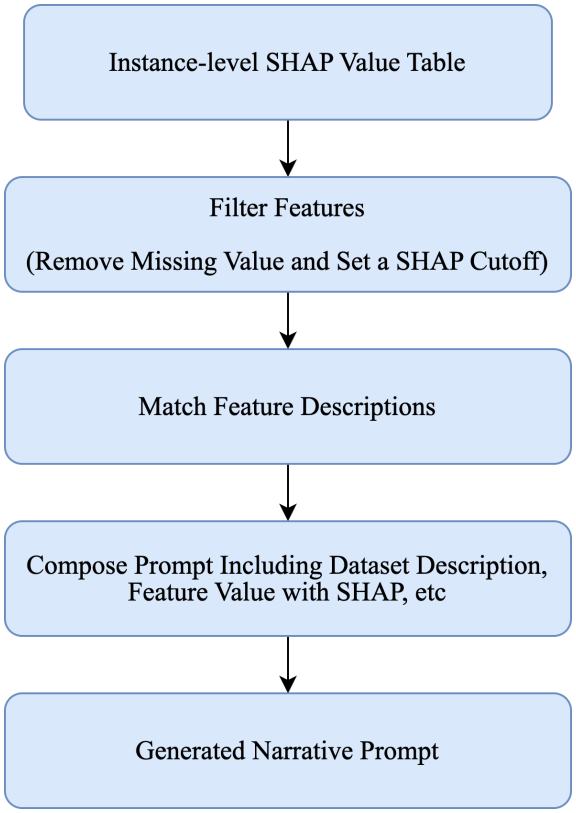

Figure 4.7: Pipeline for generating narrative prompts from SHAP-based feature attributions.

4.7.2 Prompt Design

The prompt construction process was iteratively refined across multiple versions to improve clarity, factual accuracy, and the interpretability of the generated narratives.

The initial version adopted a general format that introduced the survival model, SHAP values, and basic response requirements. However, this style often resulted in responses that lacked coherence or focused excessively on model evaluation rather than interpretability. To encourage more explanatory and storytelling-driven outputs, subsequent iterations emphasized the importance of feature attributions and explicitly discouraged assessments of model performance.

During this refinement process, we also identified instances where SHAP attributions contradicted established medical knowledge, which posed a risk of generating

misleading content. To mitigate this, explicit instructions were added to prompt the LLM to acknowledge such conflicts and interpret them in terms of what the model may have learned from its training data.

Moreover, prior research by Cruz and Lombrozo (2025) [31] suggests that while domain-specific jargon may increase perceived explanatory satisfaction, it can actually reduce comprehension for non-expert audiences. This insight informed our decision to prioritize clarity and completeness over technical terminology. Accordingly, we introduced a role-based instruction within the prompt, framing the LLM as a medical expert communicating with a lay audience. This strategy encourages the generation of more accessible and contextually appropriate narratives.

Later versions of the prompt incorporated structured medical context, clearer instructions, and narrative constraints, such as response length and tone. We experimented with various styles—clinical, descriptive, and comparative—and ultimately selected a prompt format that balances technical rigor with accessibility for non-expert audiences.

4.7.3 Local Deployment of LLMs

To ensure compliance with data privacy regulations and avoid transmitting sensitive clinical information to external servers, all large language model (LLM) inference was conducted locally. We utilized the Ollama framework to host and interact with a variety of open-source LLMs, supporting both lightweight (8–14B parameters) and larger-scale models. Initial development and testing were performed on a MacBook Pro (Apple M3 Pro), which provided sufficient capacity for running smaller models. For experiments involving more computationally demanding LLMs, such as 34B and 70B variants, inference was transitioned to a high-performance computing (HPC) cluster equipped with modern GPUs.

The software environments on both local and cluster setups were aligned by

maintaining consistent versions of all dependencies, including PyTorch, Transformers, and supporting libraries. This ensured reproducibility and compatibility across different deployment platforms.

This local deployment strategy enabled secure, controlled, and scalable evaluation of different LLMs for narrative generation without compromising patient confidentiality.

Ollama provides a lightweight and containerized environment for running language models such as DeepSeek or LLaMA locally with minimal setup. It supports efficient on-device inference and was particularly suited for iterative prompt tuning and model response inspection.

For large-scale generation tasks, we used vLLM, an open-source framework optimized for fast and memory-efficient inference of transformer-based LLMs. The vLLM deployment leveraged GPU acceleration and supported parallel request batching, which was crucial for handling multiple patient-level narratives efficiently.

5 Result and Discussion

5.1 Survival Model Performance

The predictive performance of the Random Survival Forest (RSF) was assessed using repeated 10-fold cross-validation (3 repetitions). The Cox proportional hazards model was used as a baseline for comparison. The results of model performance are in the table:

Metric	**Random Survival Forest**	**CoxPH Survival Analysis**
Mean C-index	0.6634	0.6084
Mean iBS	0.0841	0.1232

Table 5.1: Model performance of RSF and Cox Model using repeated 10-fold cross-validation (3 repeats).

A C-index above 0.65 and an iBS below 0.1 are generally considered indicative of good model performance in survival analysis. Based on these criteria, the RSF model demonstrates superior predictive ability compared to the Cox model, and is therefore selected as the preferred survival model.

Using the optimal missing value threshold and model hyperparameter settings identified through grid search, we further evaluated the RSF model separately on two different outcome events (OSA and EFSA). The performance metrics are summarized in Table 5.2.

Metric	**OSA**	**EFSA**
Mean C-index	0.6996	0.6600
Mean iBS	0.0812	0.1212

Table 5.2: Model performance of RSF using repeated 10-fold cross-validation (3 repeats).

These results suggest that a lower censoring rate does not necessarily lead to improved model performance. Other factors, such as event heterogeneity, the informativeness of available features, and the suitability of model assumptions, may also substantially influence predictive accuracy.

5.2 Prediction Time-point Selection

The Random Survival Forest (RSF) model outputs an estimated survival function for each individual. However, such function-valued predictions can be challenging to interpret or communicate in a clinically meaningful way. To address this, we derived a single-point summary—the predicted survival time—from each survival function, enabling straightforward comparison with the observed survival time.

We considered two strategies for deriving this point prediction: the **expected survival time** (i.e., the mean of the survival distribution) and the **median survival time** (i.e., the time at which the predicted survival probability falls to 0.5). Given the right-skewed nature of survival data, expected survival time can be heavily influenced by long tails, leading to potentially unstable or overly optimistic predictions. In contrast, the median survival time tends to be more robust and interpretable.

To empirically support this choice, we plotted the expected and median predicted survival times against actual survival times using uncensored samples (Figure 5.1). The median-based predictions exhibited a stronger alignment with observed values.

This was further confirmed by computing the coefficient of determination (R^2), where the median survival time yielded a score of -0.182, compared to -0.775 for the expected survival time. While both scores are negative—likely due to the effects of censoring—the median prediction still shows a relatively closer alignment to the actual survival times.

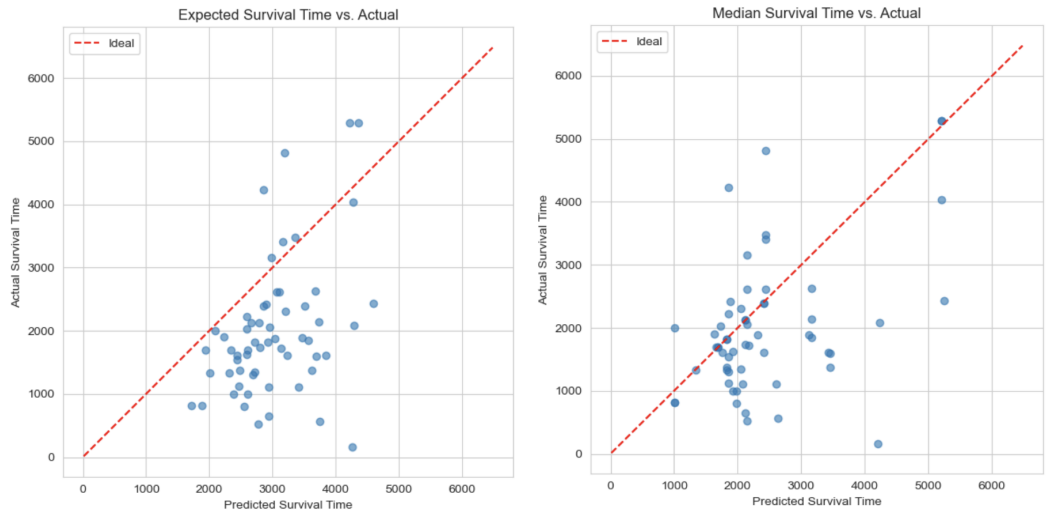

Figure 5.1: Comparison between expected survival time and median survival time against actual observed survival times. The red dashed line denotes the ideal diagonal where predicted = actual.

Based on these findings, we chose to use the **median survival time** as the primary point prediction for downstream evaluation and interpretation.

5.3 SHAP-Based Model Interpretability

To interpret the model's behavior and identify which clinical features influenced its predictions, we aggregated the instance-level SHAP values computed at each patient's median predicted survival time. By averaging the absolute SHAP values across all patients, we obtained a global view of feature contribution magnitudes. This reflects how strongly each variable contributed, on average, to model-predicted survival outcomes. The summary in Figure 5.2 highlights features with the highest

overall impact, serving as a basis for global interpretability.

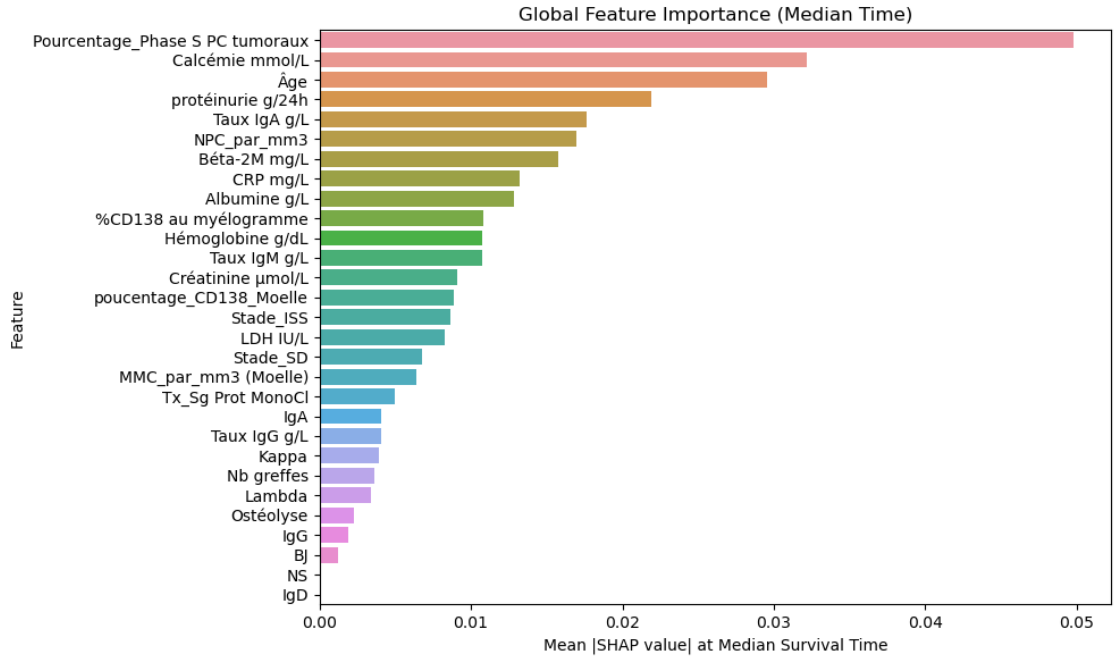

Figure 5.2: Global feature attribution based on mean absolute SHAP values computed at each patient's median predicted survival time. Features with higher values contributed more strongly to model survival predictions across the cohort.

While global SHAP analysis reveals which features were most influential across the entire cohort, it does not explain how those features contributed to individual predictions. To address this, we further analyzed local SHAP value distributions at the patient level.

We selected one sample and visualized the SHAP value profile at the median survival times. The visualization illustrates how the model attributed importance to each clinical feature for a specific patient, reflecting personalized decision rationales.

In each case, features with large positive SHAP values contributed to increased predicted survival, while negative SHAP values indicated factors associated with decreased survival. This patient-level interpretability is particularly valuable in clinical contexts, where individualized explanation is essential for transparency and trust.

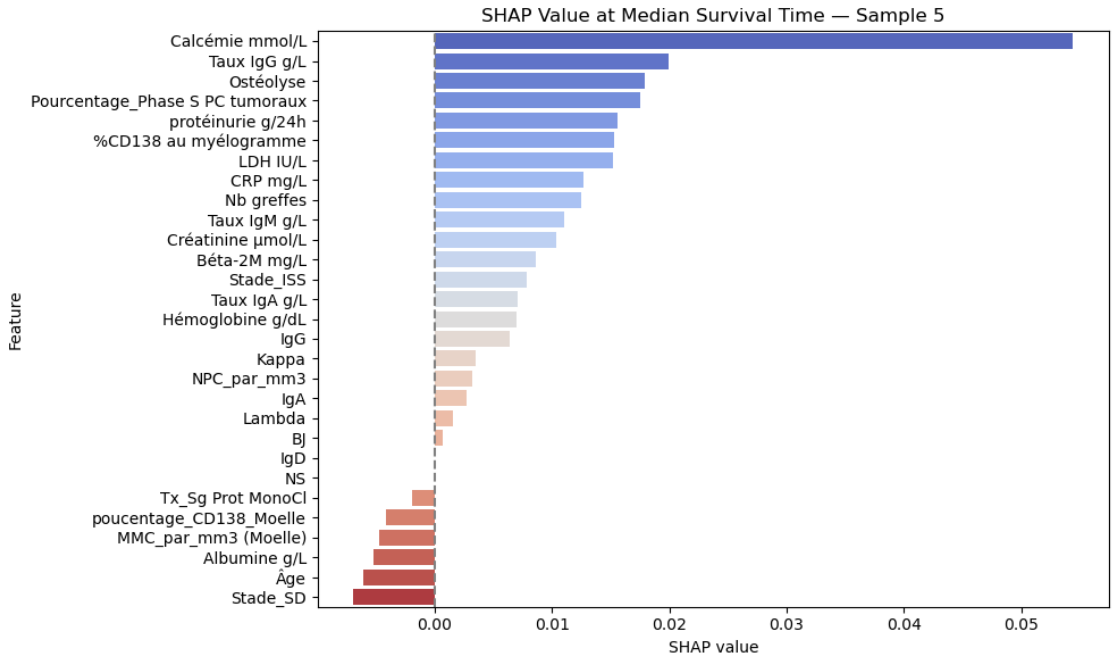

Figure 5.3: Local SHAP values for Sample 5 at the predicted median survival time.

5.4 Narrative Explanation

Building upon the local SHAP values described in the previous subsection, we generated natural language rationales to explain individual predictions. For each patient, a structured prompt was constructed based on features ranked by their absolute SHAP values at the predicted median survival time, using a cutoff threshold (default = 0.01) to retain only the most influential variables. The prompt is designed based on the research[4] and after several rounds of prompt optimization, we found that prompts containing detailed descriptions of medical features and using rankings based on the absolute SHAP values—rather than the raw SHAP values—led to more coherent and informative outputs from the language model.

The finalized prompt format is shown below:

Final Prompt Template

You are a data scientist explaining an AI model's prediction using computed SHAP values. You need to interpret the feature attributions for a medical expert and explain what positive and negative SHAP values represent.

Here's the background information for the patient case we are analyzing:
An AI survival model was used to predict a dataset on patients diagnosed with multiple myeloma, containing diagnostic information, biomarkers, treatment history, and clinical outcomes. The input features of the data include data about the complete clinical profile of a patient with multiple myeloma. The target variables include the event indicator (whether the event occurred or was censored) and the time to the event, representing the predicted duration until death from multiple myeloma.

The predicted result and the actual target variables of a certain instance are shown below: The AI model was trained to predict median survival time and the model predicted a median survival time of {predicted} days. This patient's actual survival time was {actual} days (the event occurred). Please ensure your explanation focuses solely on **why the model made this prediction** and **does not** judge or estimate the model's performance in this case.

The provided SHAP table was generated to help us understand this outcome. It includes every feature along with its value for that instance, and the SHAP value assigned to it. For this explanation, a positive SHAP value indicates that a given variable has increased the survival possibility (a more favorable outcome), while a negative value indicates a decrease (a less favorable outcome).

Final Prompt Template

The goal of SHAP is to explain the prediction of an instance by computing the contribution of each feature to the prediction. The SHAP explanation method computes Shapley values from coalitional game theory. The feature values of a data instance act as players in a coalition. Shapley values tell us how to fairly distribute the "payout" (= the prediction) among the features. If the interpretation of a feature's SHAP value appears to conflict with established medical facts, provide a concise explanation that emphasizes *what the model learned from its specific data patterns*. Your explanation should be accurate to the model's behavior and convincing, without misleading the audience by distorting medical facts.

Now, your task is to come up with a plausible, fluent, and correct story to explain why the model predicted this outcome. Focus on the features with the highest absolute SHAP values, and explain what positive and negative SHAP values indicate. In your story, try to explain the most important feature values and potential interactions that fit the narrative. There is no need to enumerate individual features outside of the story. Conclude with a short summary of why this predicted survival time may have resulted. Limit your answer to 10-14 sentences.

{Features with Value and SHAP Value Table}

{Corresponding Feature Description}

5.5 Comparison of LLM Responses

In evaluating the generated narratives from various locally deployed LLMs, we observed notable differences across three key dimensions: factual consistency with

SHAP-based feature attributions, logical reasoning grounded in feature descriptions, and narrative fluency and clarity. Among all evaluated models, **DeepSeek-R1-Distill-Llama-70B** demonstrated the most balanced performance, exhibiting strong alignment with SHAP values, clinically coherent reasoning, and fluent storytelling. Notably, it is also the largest model we deployed locally on the computational cluster, with the highest number of parameters among all tested LLMs.

Model	SHAP Consistency	Model Reasoning	Narrative Quality	Notes
DeepSeek-R1-Distill-Llama-70B	✓	✓	✓	Fluent and well-aligned with SHAP attribution
Yi-34B-Chat	✗	✓	✓	Misinterprets SHAP direction despite valid reasoning
Qwen1.5-32B-Chat	✗	•	✓	Coherent narrative, but lacks logical consistency
Mixtral-8x7B	✗	✗	✗	Contradictory output, weak justifications
DeepSeek-r1:14b	✗	✗	✗	Low factual accuracy and interpretability

Table 5.3: Qualitative evaluation of local LLMs across three dimensions: SHAP factual consistency, medical reasoning, and narrative quality.

Here is a representative example generated by DeepSeek-R1-Distill-Llama-70B. To protect patient privacy, detailed numerical values and figures are omitted. After being provided with clinical feature descriptions, the LLM is better able to explain the model's decision-making rationale and address potential contradictions between SHAP-based reasoning and established medical knowledge.

Response Example

The AI model predicted a median survival time of 1660.7 days for the patient, considering several key factors that negatively impacted the prognosis. According to the SHAP analysis, the model's prediction was influenced by features with negative contributions, indicating factors that decrease survival possibility.

1. **Hypercalcemia (Calcémie mmol/L: 4.89)**: Elevated calcium levels suggest bone resorption and increased tumor activity, contributing negatively

Response Example

to survival with a SHAP value of -0.1555.

2. **IgA Subtype (IgA: 1.0)**: The presence of the IgA subtype may indicate specific biological features associated with a worse prognosis, contributing a SHAP of -0.0327.

3. **High S Phase Percentage (3.0%)**: Indicating rapid tumor cell proliferation, this factor suggests aggressive disease and contributes a SHAP of -0.0293.

4.**Hemoglobin (15.4 g/dL)**: Although within normal range, the model associates higher hemoglobin with worse outcomes in this dataset, contributing a SHAP of -0.0197.

5. **Elevated IgA Levels (32.0 g/L)**: High IgA levels are indicative of the IgA myeloma subtype, contributing a SHAP of -0.0195.

6. **Low Albumin (35.0 g/L)**: Reflecting possible inflammation or malnutrition, this contributes a SHAP of -0.0182.

These factors collectively pulled the prediction downward, yet the model still estimated a longer survival time than the observed 1359 days. This discrepancy suggests the model may have considered other positive factors not listed or a high baseline survival estimate. The prediction reflects the model's weighting of these negative factors against other potential positive influences, resulting in a longer predicted survival despite adverse prognostic indicators.

6 Conclusion

This study proposed a hybrid framework that combines survival model predictions, structured prompt generation, and locally deployed large language models (LLMs) to produce narrative explanations for multiple myeloma patients. Using a Random Survival Forest (RSF), we predicted median survival times and extracted time-dependent SHAP values at that specific time point via SurvSHAP(t), enabling individualized interpretability. These SHAP values were used to construct structured prompts, which were then processed by LLMs to generate human-readable narratives. The resulting explanations offer valuable insights into feature attributions for the domain experts.

Compared to Cox model, the RSF model achieved more ideal and precise prediction in the survival analysis with higher dimensional dataset in this case. While the RSF model establishes a reasonable predictive baseline, there remains room for improvement. For example, in some instances, older age was incorrectly attributed as a positive survival factor—contradicting established clinical understanding. This suggests potential limitations of the model in capturing complex feature interactions or cohort-specific biases. Future work may explore deep learning-based survival models that can better account for such non-linear dependencies.

Regarding the LLM-based narrative generation, prompts were carefully crafted to minimize ambiguity and support domain-specific reasoning. However, the language models employed in this study exhibited limited understanding of medical and

biological concepts. To address this, we explicitly incorporated background information—such as normal ranges and the clinical significance of abnormal values—into the prompts to improve the interpretability of the generated explanations. In future work, the description of biomarkers should be contextualized based on the characteristics of the patient cohort, rather than relying on normal ranges derived from healthy individuals, to enhance the relevance and interpretability of explanations in the context of multiple myeloma.

Due to computational and resource constraints, our experiments primarily focused on small- to medium-sized LLMs deployed locally. To compensate for their limited domain knowledge, we enriched the prompts with detailed feature descriptions and contextual cues. In future work, the use of more powerful, larger-scale models may further enhance both the reasoning accuracy and narrative quality in clinical explanation tasks.

Finally, while our evaluation revealed notable differences in model performance, the assessment primarily focused on quantitative metrics and model-generated outputs. The narrative explanations produced by the LLMs were not yet been evaluated by a broader audience. Consequently, the interpretability and trustworthiness of these narratives remain to be systematically validated. Future work will incorporate expert-centered evaluation and qualitative analysis to better assess the clarity, relevance, and clinical utility of the generated explanations.

Overall, this study underscores the importance of combining model interpretability techniques with generative AI to improve transparency in clinical decision-making. Our framework may serve as a step toward more explainable AI applications in healthcare.

7 Declaration of AI Usage in the Thesis

This project centers on the use of large language models (LLMs) for generating individualized narrative explanations based on feature attributions from SHAP value. In particular, locally deployed LLMs were integrated into the experimental pipeline to generate sample-wise rationales, interpreting model predictions in natural language. These explanations are a core part of the final outputs.

In the research process, additional AI tools were employed to assist with understanding related work, refining code, and improving writing clarity. Besides Google Scholar, Consensus served as a supporting tool to help identify relevant papers and extract key insights. In addition, tools such as Gemini and NotebookLM were used to interpret complex methods and summarize scientific contents from the collected literature.

For code development, ChatGPT provided suggestions on implementation techniques and guidances. It also served as an interactive assistant to help clarify the overall project logic. Claude was used for efficient debugging and code refinement.

During the thesis writing stage, ChatGPT was occasionally used to improve sentence fluency, check for grammatical issues, and rephrase my original text where appropriate. All conceptual contributions, technical implementation, and final manuscript structuring remain my own.

Overall, while AI tools supported various aspects of the project, their roles were limited to assistance in understanding, coding, and writing. The LLM-generated narratives in the results are part of the core experimental design and output of this work.

References

[1] T. G. Clark, M. J. Bradburn, S. B. Love, and D. G. Altman, "Survival Analysis Part I: Basic concepts and first analyses", *British Journal of Cancer*, vol. 89, no. 2, pp. 232–238, Jul. 2003. DOI: 10.1038/sj.bjc.6601118.

[2] S. M. Lundberg, G. Erion, H. Chen, *et al.*, "From local explanations to global understanding with explainable AI for trees", *Nature Machine Intelligence*, vol. 2, no. 1, pp. 56–67, Jan. 2020. DOI: 10.1038/s42256-019-0138-9.

[3] M. F. Dahlstrom, "Using narratives and storytelling to communicate science with nonexpert audiences", *Proceedings of the National Academy of Sciences*, vol. 111, no. supplement_4, pp. 13 614–13 620, Sep. 2014. DOI: 10.1073/pnas.1320645111.

[4] D. Martens, J. Hinns, C. Dams, M. Vergouwen, and T. Evgeniou, "Tell me a story! Narrative-driven XAI with large language models", *Decision Support Systems*, vol. 191, p. 114 402, 2025. DOI: 10.1016/j.dss.2025.114402.

[5] A. J. Turkson, F. Ayiah-Mensah, and V. Nimoh, "Handling Censoring and Censored Data in Survival Analysis: A Standalone Systematic Literature Review", *International Journal of Mathematics and Mathematical Sciences*, vol. 2021, N. Tang, Ed., pp. 1–16, Sep. 2021. DOI: 10.1155/2021/9307475.

[6] D. G. Altman and J. M. Bland, "Time to event (survival) data", *BMJ (Clinical research ed.)*, vol. 317, no. 7156, pp. 468–469, 1998. DOI: 10.1136/bmj.317.7156.468.

[7] J. Báskay, T. Mezei, P. Banczerowski, A. Horváth, T. Joó, and P. Pollner, "Censoring Sensitivity Analysis for Benchmarking Survival Machine Learning Methods", *Sci*, vol. 7, no. 1, p. 18, Feb. 2025. DOI: 10.3390/sci7010018.

[8] S. Germer, C. Rudolph, L. Labohm, *et al.*, "Survival analysis for lung cancer patients: A comparison of Cox regression and machine learning models", *International Journal of Medical Informatics*, vol. 191, p. 105 607, Nov. 2024. DOI: 10.1016/j.ijmedinf.2024.105607.

[9] H. Wang and G. Li, "A Selective Review on Random Survival Forests for High Dimensional Data", *Quantitative Bio-Science*, vol. 36, no. 2, pp. 85–96, Nov. 2017. DOI: 10.22283/QBS.2017.36.2.85.

[10] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, "Random survival forests", *The Annals of Applied Statistics*, vol. 2, no. 3, pp. 841–860, 2008. DOI: 10.1214/08-AOAS169.

[11] D. M. Senevirathne, S.-I. Yang, C. Brandeis, and D. G. Hodges, "Predicting time-to-harvest in mixed-species forests using a random survival forest algorithm", *Forest Ecosystems*, vol. 11, p. 100 236, 2024. DOI: 10.1016/j.fecs.2024.100236.

[12] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions", in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, *et al.*, Eds., vol. 30, Curran Associates, Inc., 2017.

[13] *Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (general data protection regulation)*, https://eur-lex.europa.eu/eli/reg/2016/679/oj, Official Journal of the European Union, L119, 4 May 2016, pp. 1–88, 2016.

[14] Z. Li, "Extracting spatial effects from machine learning model using local interpretation method: An example of SHAP and XGBoost", *Computers, Environment and Urban Systems*, vol. 96, p. 101 845, Sep. 2022. DOI: 10.1016/j.compenvurbsys.2022.101845.

[15] A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller, and W. Samek, Eds., *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers* (Lecture Notes in Computer Science). Cham: Springer International Publishing, 2022, vol. 13200. DOI: 10.1007/978-3-031-04083-2.

[16] M. Rezaei, L. Tapak, M. Alimohammadian, A. Sadjadi, and M. Yaseri, "Review of Random Survival Forest method", *Journal of Biostatistics and Epidemiology*, Nov. 2020. DOI: 10.18502/jbe.v6i1.4760.

[17] S. M. Lundberg, G. G. Erion, and S.-I. Lee, *Consistent Individualized Feature Attribution for Tree Ensembles*, Mar. 2019. DOI: 10.48550/arXiv.1802.03888.

[18] M. A. K. Raiaan, M. S. H. Mukta, K. Fatema, *et al.*, "A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges", *IEEE Access*, vol. 12, pp. 26 839–26 874, 2024. DOI: 10.1109/ACCESS.2024.3365742.

[19] H. Zhao, H. Chen, F. Yang, *et al.*, "Explainability for Large Language Models: A Survey", *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 2, pp. 1–38, Apr. 2024. DOI: 10.1145/3639372.

[20] *Health insurance portability and accountability act of 1996 (hipaa)*, https://www.hhs.gov/hipaa/for-professionals/privacy/index.html, Public Law 104-191, 1996.

[21] *Regulation (eu) 2024/1689 of the european parliament and of the council of 12 july 2024 laying down harmonised rules on artificial intelligence (artificial intelligence act)*, https://eur-lex.europa.eu/eli/reg/2024/1689/oj, Official Journal of the European Union, L, 12 July 2024, 2024.

[22] International Myeloma Foundation, *Durie-salmon staging system*, Accessed: 2025-05-29, 2021. [Online]. Available: https://www.myeloma.org/durie-salmon-staging.

[23] International Myeloma Foundation. "International staging system (iss) & revised iss (r-iss)". Accessed: 2025-05-24. (2023), [Online]. Available: https://www.myeloma.org/international-staging-system-iss-reivised-iss-r-iss.

[24] E. W. Steyerberg, A. J. Vickers, N. R. Cook, *et al.*, "Assessing the Performance of Prediction Models: A Framework for Traditional and Novel Measures", *Epidemiology*, vol. 21, no. 1, pp. 128–138, Jan. 2010. DOI: 10.1097/ede.0b013e3181c30fb2.

[25] E. Graf, C. Schmoor, W. Sauerbrei, and M. Schumacher, "Assessment and comparison of prognostic classification schemes for survival data", *Statistics in Medicine*, vol. 18, no. 17-18, pp. 2529–2545, Sep. 1999. DOI: 10.1002/(sici)1097-0258(19990915/30)18:17/18<2529::aid-sim274>3.0.co;2-5.

[26] S. Y. Park, J. E. Park, H. Kim, and S. H. Park, "Review of Statistical Methods for Evaluating the Performance of Survival or Other Time-to-Event Prediction Models (from Conventional to Deep Learning Approaches)", *Korean Journal of Radiology*, vol. 22, no. 10, p. 1697, 2021. DOI: 10.3348/kjr.2021.0223.

[27] M. Krzyziński, M. Spytek, H. Baniecki, and P. Biecek, "SurvSHAP(t): Time-dependent explanations of machine learning survival models", *Knowledge-Based*

Systems, vol. 262, p. 110 234, Feb. 2023. DOI: 10.1016/j.knosys.2022.110234.

[28] Z. Ying, S. H. Jung, and L. J. Wei, "Survival analysis with median regression models", *Journal of the American Statistical Association*, vol. 90, no. 429, pp. 178–184, 1995. DOI: 10.1080/01621459.1995.10476500.

[29] *REID*, "Estimating the median survival time", *Biometrika*, vol. 68, no. 3, pp. 601–608, Dec. 1981. DOI: 10.1093/biomet/68.3.601.

[30] O. Ben-Aharon, R. Magnezi, M. Leshno, and D. A. Goldstein, "Median survival or mean survival: Which measure is the most appropriate for patients, physicians, and policymakers?", *The Oncologist*, vol. 24, no. 11, pp. 1469–1478, Jul. 2019. DOI: 10.1634/theoncologist.2019-0175.

[31] F. Cruz and T. Lombrozo, "How laypeople evaluate scientific explanations containing jargon", *Nature Human Behaviour*, Jun. 2025. DOI: 10.1038/s41562-025-02227-0.