Universidad Politécnica
de Madrid

Escuela Técnica Superior de Ingenieros Informáticos

Master in Health and Medical Data Analytics

Master Thesis

Machine Learning models for predicting Cure or Relapse in Post Kala Azar Dermal Leishmaniasis (PKDL)

Author: María del Mar Moreno Ocaña
Tutors: Miguel García Remesal and Eugenia Carrillo Gallego

Madrid, July - 2025

This Master Thesis has been deposited in ETSI Informáticos de la Universidad Politécnica de Madrid.

Master Thesis
Master in Health and Medical Data Analytics

Title: Machine Learning models for predicting Cure or Relapse in Post Kala Azar Dermal Leishmaniasis (PKDL)

July - 2025

Author: María del Mar Moreno Ocaña
Tutors: Miguel García Remesal
Departamento de Inteligencia Artificial
ETSI Informáticos
Universidad Politécnica de Madrid

Eugenia Carrillo Gallego
Doctora en Microbiología y Parasitología
Científica Titular de OPI
Centro Nacional de Microbiología
Instituto de Salud Carlos III

Acknowledgements

First of all, I would like to thank Eugenia and Miguel for giving me the opportunity to carry out this project and for always being so willing to answer my questions and offer valuable advice.

I also want to thank my parents — without their support, I wouldn't be here doing this work, and I'm deeply grateful for their patience and encouragement through all my moments of stress.

Finally, I want to thank all the wonderful friends I've made during this Master's program, who have become like family — and especially David, for always being there and helping me with everything.

Table of Contents

Abstract

Post-Kala-Azar Dermal Leishmaniasis (PKDL) is a neglected tropical disease that emerges in some patients after treatment for visceral leishmaniasis. Its clinical management is challenging due to complex immune responses and the potential for relapse. This Master Thesis aims to explore patterns in PKDL patient data through both unsupervised and supervised machine learning methods. Initially focused on exploratory clustering, the study evolved to include classification models after identifying patients who experienced relapse or required rescue treatment.

A curated set of clinical, biochemical, hematological, and immunological biomarkers was analyzed, with particular emphasis on cytokines, organ function indicators, and parasite load. Two new outcome variables—*Rescue* and *Relapse*—were engineered to better capture patient trajectories. Through this data-driven approach, the project seeks to identify predictive factors of treatment failure and deepen the understanding of disease progression in PKDL. The findings are interpreted in light of recent literature and aim to support more effective treatment strategies.

Chapter 1

Introduction

This chapter provides an overview of the clinical context and motivation behind the study, defines the main problem and objectives, and outlines the expected outcomes and impact. It also explains the use of data-driven approaches to study treatment response and relapse in PKDL, and presents the structure and planning of the project.

1.1. Motivation

Post-kala-azar dermal leishmaniasis (PKDL) is a neglected tropical disease that disproportionately affects vulnerable populations, especially in endemic regions such as Sudan. It often appears after apparent clinical cure from visceral leishmaniasis (VL), and while treatment is available, a significant number of patients experience relapse. These relapses not only hinder the individual recovery process but also represent a major obstacle for public health efforts aimed at eradicating the disease. [1]

Despite advances in our understanding of the immunological and molecular responses involved in PKDL, there is still no reliable way to predict which patients will relapse after treatment. This underscores the urgent need to identify robust biomarkers that can distinguish between patients likely to achieve sustained cure and those at risk of relapse. Such biomarkers could significantly improve treatment decisions, reduce unnecessary exposure to second-line therapies, and support the long-term control of leishmaniasis.

With a background rooted in health and a strong personal motivation to apply technology to solve real-world medical challenges, I see this project as an opportunity to contribute meaningfully to global health. Leveraging un-

supervised machine learning to identify clinically relevant biomarker profiles aligns with this goal and offers the potential to enhance our understanding of PKDL treatment response in a data-driven and scalable way.

1.2. Problem Statement

PKDL is a dermal manifestation that occurs following apparent clinical cure of visceral leishmaniasis (VL). Despite treatment, a proportion of PKDL patients experience disease relapse, which presents major challenges for disease elimination efforts. Identifying robust biomarkers that can predict treatment outcomes, particularly relapse, remains a significant unmet need. [2]

Although several immunological and molecular mechanisms underlying PKDL have been explored, current treatments do not guarantee long-term cure for all patients. The variability in relapse rates between treatment regimens suggests that the underlying immune response to therapy may play a pivotal role in determining disease outcome [3]. Understanding and distinguishing the immunological profiles associated with sustained cure versus relapse could not only optimize treatment selection but also inform the development of predictive biomarkers.

Therefore, the central problem addressed in this study is the lack of reliable predictive markers of relapse in PKDL patients. Specifically, we aim to identify early biomarkers, measurable at baseline or shortly after treatment, that are predictive of long-term treatment failure or success across different therapeutic regimens.

Treatment of PKDL in Sudan has historically relied on two main regimens based on expert consensus and clinical practice guidelines:

1. **Sodium stibogluconate (SSG):** A pentavalent antimonial administered intravenously or via painful intramuscular injections daily for 60 days or more. While effective, it is associated with notable toxicity (cardiac, hepatic, pancreatic) and requires extended hospitalization, which poses logistical challenges in resource-constrained settings [4].
2. **Combination therapy with SSG and Paromomycin (PM):** Introduced to shorten treatment duration, this regimen couples SSG with the aminoglycoside PM over approximately 17 days. Clinical data from multiple East African sites (including Sudan) have demonstrated similar efficacy to prolonged SSG monotherapy, with improved safety, shorter hospital stays, and lower cost [4, 5].

1.2.0.1. 1.2.1 Differences in Immune Response Between Treatment Arms

The study by Torres et al. [6] highlights key differences in the immune response dynamics between the two treatment arms used in PKDL therapy.

Arm 1 consisted of a once-daily intramuscular injection of 20 mg/kg/day Paromomycin (PM) for 14 days combined with oral Miltefosine (MF) administered twice daily for 42 days. Patients in this arm exhibited a slower and less consistent recovery of antigen-specific IFN-γ responses. Notably, some patients who initially appeared cured later experienced clinical relapse.

In contrast, Arm 2 comprised four intravenous infusions of 5 mg/kg Liposomal Amphotericin B (LAmB) administered on days 1, 3, 5, and 7 (total dose of 20 mg/kg), along with oral MF twice daily for 28 days. This regimen was associated with a more rapid and sustained restoration of Th1-type immune responses, characterized by significantly elevated levels of IFN-γ and TNF-α, which are essential for effective leishmanicidal activity.

The immunological divergence observed between the two arms suggests that the nature and timing of immune restoration during treatment may play a decisive role in determining long-term outcomes. The absence of relapses in the LAmB/MF group (Arm 2) underscores the potential superiority of this regimen in inducing durable protective immunity.

These findings emphasize the need to deepen our understanding of the immunological factors that differentiate relapse from sustained cure in PKDL. The observation that specific immune markers—such as IFN-γ and TNF-α—are modulated differently depending on the treatment strategy, suggests the existence of predictive biomarkers that may remain undetectable using conventional statistical approaches [3, 7].

The primary objective of this Master Thesis is to identify potential biomarkers associated with PKDL cure and to elucidate factors contributing to relapse. This is achieved by analyzing 31 clinical, biochemical, hematological, and immunological parameters collected from 110 Sudanese patients diagnosed with PKDL. Data were gathered at baseline and post-treatment, with follow-up assessments over a two-year period to monitor clinical outcomes.

The study leverages both supervised and unsupervised machine learning approaches. Unsupervised techniques are employed to uncover latent structures and natural groupings within the dataset—potentially revealing immunological profiles or clinical trajectories associated with treatment response [8]. In parallel, supervised models are developed to distinguish between

relapsing and cured patients. The combination of both methodologies allows for robust prediction while uncovering complex relationships that traditional methods may overlook.

1.3. Objetives

The main objective of this Master Thesis is the identification of cure biomarkers in post-kala-azar dermal leishmaniasis (PKDL). The following are the specific objectives that will guide the development of this project.

1. **Data Preparation and Preprocessing:** The first step is to preprocess and clean the dataset, which includes 31 measured factors collected before and after treatment. This involves handling missing values, standardizing the data to ensure comparability between variables, and preparing the dataset for the application of machine learning models.
2. **Application of Unsupervised Machine Learning Methods:** After data preparation, unsupervised machine learning algorithms, such as clustering techniques, will be applied to identify groups of biomarkers that are associated with the cure of PKDL or with the risk of relapse. The goal is to identify meaningful patterns and relationships within the data that can inform treatment strategies.
3. **Application of Supervised Machine Learning Methods:** Originally, the objective was limited to unsupervised analysis due to the absence of clearly labeled outcomes. However, once relapse cases were identified and integrated into the dataset, the study was expanded to also explore supervised learning techniques. This allowed for the development of classification models aimed at predicting relapse, thereby complementing the exploratory insights obtained through unsupervised methods.
4. **Validation of Results:** The results from the machine learning models will be validated by comparing the identified biomarkers with existing clinical evidence and previously published studies on PKDL. Collaboration with experts in microbiology and clinical practice will ensure that the findings are robust and clinically relevant. This step is essential for generating new hypotheses and refining our understanding of the disease.
5. **Proposing New Therapeutic Strategies:** Based on the findings from the machine learning models, this study will propose potential biomarkers that could guide personalized treatment strategies for PKDL.

These strategies may lead to more effective therapies, tailored to the individual characteristics of patients, ultimately improving treatment outcomes and reducing the risk of relapse.

These objectives are essential for the success of the project and will enable the identification of key cure biomarkers in PKDL. This will contribute to improving treatment strategies and reducing relapses, ultimately benefiting those affected by this neglected tropical disease.

1.4. Expected Outcomes

The anticipated outcomes of this research include:

1. Identification of key biomarkers that are associated with the cure of PKDL, potentially allowing for more effective treatment strategies.
2. A preliminary classification of patients based on their risk of relapse, which could guide clinical decision-making and follow-up care.
3. Development of an analytical pipeline for identifying biomarkers in similar future studies of infectious diseases or chronic conditions.

1.5. Impact

The findings from this Master Thesis could significantly enhance the understanding of the immune and clinical factors that influence PKDL treatment outcomes. By identifying reliable biomarkers of cure and relapse, this research lays the groundwork for the development of personalized medicine strategies that can be applied to PKDL. Furthermore, these insights could inform future studies and contribute to the broader field of infectious disease research, particularly in diseases with complex immune responses like leishmaniasis.

This research also holds the potential to improve patient outcomes by reducing relapse rates and ensuring more targeted interventions for patients suffering from PKDL. Ultimately, the identification of these biomarkers could lead to more effective, individualized treatments that address the underlying causes of relapse and enhance the overall cure rate of this debilitating disease.

1.6. Project Management

To ensure efficient management and tracking of progress, several tools were used to plan and organize the timeline through sprints. Each sprint groups a coherent set of related tasks, allowing for focused execution and easier monitoring. The Gantt chart in Figure 1.1 visually illustrates the task distribution, duration, and dependencies across the project lifecycle.

Sprint 0: Follow-up (6 months)
Throughout the project, 18 follow-up sessions were scheduled. These 1-day check-ins served to monitor progress, address questions, realign objectives, and maintain accountability. Their regular cadence was essential to ensuring consistent momentum.

Sprint 1: Initial Familiarization (March)
This sprint aimed to build a foundational understanding through parallel tasks:

1. **Clinical and Dataset Familiarization:** Analyzing the clinical context and understanding the structure and content of the dataset.
2. **Literature Review:** Reviewing relevant scientific studies to identify key findings, gaps, and methodological insights.
3. **Writing Report:** Early documentation efforts began to capture foundational insights and structure.

Sprint 2: Exploration Preprocessing (March – April)
This sprint focused on preparing the data for analysis:

1. **Variable Characterization:** Assessing each variable's type, distribution, and clinical meaning.
2. **Data Preprocessing:** Cleaning, transforming, and handling missing data to ensure quality input.
3. **Exploratory Data Analysis (EDA):** Investigating trends, correlations, and preliminary patterns.

Sprint 3: Validation Results (April – May)
This phase involved assessing the robustness of findings and assembling them clearly:

1. **Results Compilation:** Organizing visualizations and outcomes for reporting.

2. **Validation and Interpretation:** Evaluating the outputs' clinical and statistical validity.

Sprint 4: Modeling and Biomarker Analysis (May – July)
Analytical development took place in this sprint:

1. **Model Comparison:** Testing different unsupervised and supervised approaches for clustering and structure discovery.
2. **Biomarker Analysis:** Identifying key variables with high influence on PKDL relapse.

Sprint 5: Finalization and Defense (June – July)
Wrapping up the work included:

1. **Revisions and Defense Preparation:** Refining the report, creating the presentation and preparing for oral defense.

The Gantt chart below (Figure 1.1) provided a clear roadmap for the execution of each sprint. Organizing the project in this manner enabled structured progress, facilitated regular feedback, and allowed flexibility where needed.

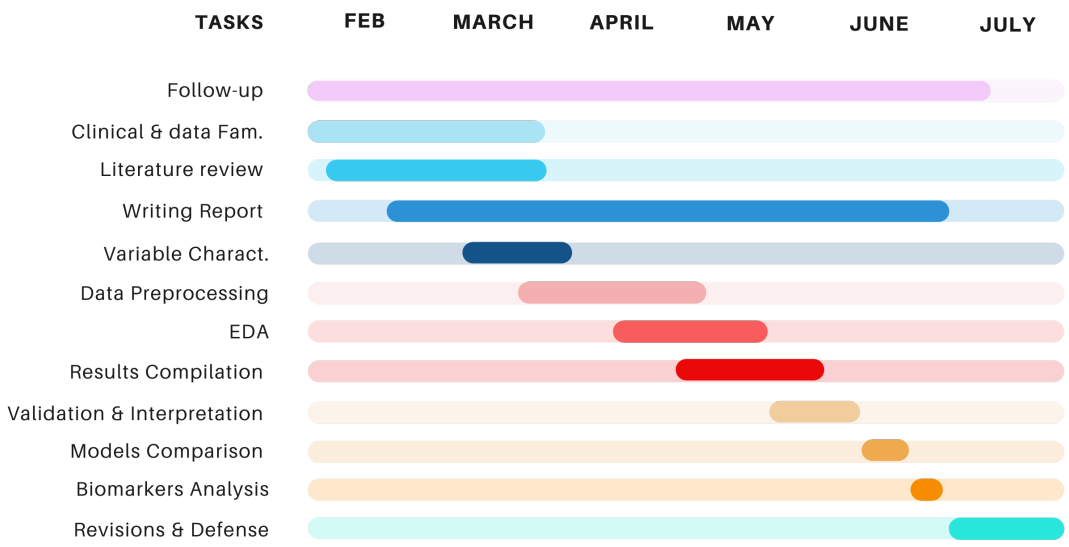


Figure 1.1: Gantt chart representing the project's tasks distributed across thematic sprints.

1.7. Development Environment and Tools

The analytical development was conducted using **Visual Studio Code** as the primary integrated development environment (IDE) [9], in conjunction with **Python Notebooks**. This environment offered the flexibility required for iterative analysis and model development, along with robust support for visualization, documentation, and reproducibility [10].

Visual Studio Code provided an efficient and extensible platform with key features such as interactive Python execution, integrated terminal, and compatibility with version control systems [9]. The use of Python Notebooks facilitated modular coding, rapid experimentation, and clear tracking of analytical workflows.

1.7.1. Core Libraries for Data Handling and Manipulation

The foundational data analysis operations were conducted using:

- **pandas**: Used extensively for data manipulation, cleaning, and tabular transformations. Its DataFrame structure allowed efficient handling of large datasets and facilitated preprocessing steps.
- **numpy**: Provided support for numerical operations and array manipulation, ensuring efficient computation throughout the pipeline.

1.7.2. Data Visualization Libraries

To support data exploration and results communication, the following libraries were employed:

- **matplotlib**: Served as the primary tool for generating static plots, histograms, and scatter plots [11].
- **seaborn**: Built on top of matplotlib, seaborn allowed enhanced statistical plotting and improved the aesthetic quality of the visualizations [12].
- **mpl_toolkits.mplot3d**: Enabled three-dimensional visualizations, particularly relevant for cluster analysis and dimensionality reduction results.

1.7.3. Preprocessing and Feature Engineering

Several libraries were used to prepare the data for modeling:

- **scikit-learn** [13]:
  - `StandardScaler`: Used to standardize numerical variables.
  - `OneHotEncoder`: Applied to transform categorical variables into binary indicators.
  - `train_test_split`: Used for dividing the data into training and testing sets.
- **re**: Python's regular expressions module, used for text preprocessing and pattern detection.
- **imblearn.over_sampling.SMOTE**: Applied to balance class distribution using synthetic sampling [14].
- `Pipeline`: Enabled streamlined chaining of preprocessing steps and model fitting in a single object.

1.7.4. Dimensionality Reduction and Clustering

For unsupervised learning tasks, multiple algorithms were utilized:

- **Principal Component Analysis (PCA)**: Reduced data dimensionality while retaining variance.
- **UMAP**: A non-linear dimensionality reduction technique used for structure-preserving projections in clustering contexts.
- **NearestNeighbors**: Used to compute distances between data points, particularly for density-based clustering algorithms like DBSCAN and for UMAP graph construction.
- **KMeans**, **DBSCAN**, **HDBSCAN**: These clustering algorithms enabled identification of patterns and subgroups in the dataset.
- `silhouette_score`, `adjusted_rand_score`, `homogeneity_score`: Used to evaluate clustering quality.

1.7.5. Statistical Analysis

Statistical testing was performed using:

- **scipy.stats**:
  - `ttest_ind`, `mannwhitneyu`, `chi2_contingency`, `shapiro`: Employed to test group differences and distributions.

- **statsmodels** [15]:
  - `multipletests`: Applied to correct for multiple testing.
  - `variance_inflation_factor`: Used to assess multicollinearity.
  - `Logit`, `api`: Provided additional statistical modeling capabilities.

1.7.6. Supervised Learning Algorithms

A broad range of classification algorithms were implemented using:

- **scikit-learn** [13]:
  - `LogisticRegression`, `RandomForestClassifier`, `SVC`, `OneClassSVM`, `IsolationForest`: Enabled training of baseline and advanced classifiers.
  - `classification_report`, `roc_auc_score`, `confusion_matrix`, `roc_curve`, `auc`: Used for model evaluation.
- **XGBoost**:
  - `XGBClassifier`: Applied for high-performance gradient boosting classification tasks [16].
- **LightGBM** and **CatBoost**:
  - Provided efficient implementations for gradient boosting, optimized for performance on structured data.

1.7.7. Model Interpretation and Feature Selection

To enhance model transparency and interpretability:

- **SHAP**: Used for feature contribution analysis across tree-based models and to interpret individual predictions [17].
- **sklearn.feature_selection.RFE**: Recursive Feature Elimination was applied to identify the most relevant features for model performance.

1.7.8. Model Persistence

To save and load trained models efficiently:

- **joblib**: Used to serialize trained models and pipelines for future reuse and deployment [18].

Chapter 2

State of art

This chapter provides an in-depth overview of the structure, composition, and clinical significance of the variables included in the dataset used for this study. The goal is to offer a solid foundation for the analyses presented in later chapters by clearly describing both original and derived variables, as well as the rationale behind the selection of key biomarkers.

2.1. Initial Steps and Familiarization with the Problem

To begin my work on this Master Thesis, I first undertook a thorough process to understand the context and challenges associated with Post-Kala-Azar Dermal Leishmaniasis (PKDL). To better understand the context, I reviewed several related research articles, which are detailed in the following section. Through this review, it became clear that PKDL is a neglected tropical disease that predominantly affects individuals in extremely low-income settings and, as such, remains under-researched. The condition is chronic and disfiguring, and its treatment and follow-up are logistically complex. In the dataset I worked with, 110 patients were treated and monitored, completing their treatment by day 42, with cutaneous symptoms observed from day 0. A 24-month follow-up was conducted, and 5 patients experienced relapse in one of the treatment arms.

One of my first steps was to carry out an exhaustive study of the dataset variables, focusing on understanding their meaning, types (categorical, numerical, temporal), and clinical relevance. This preliminary work was crucial

to identify possible patterns or relationships between variables and relapse status, and to later guide the data analysis and modeling process.

2.2. Literature Review

To gain a deeper understanding of Post-Kala-Azar Dermal Leishmaniasis (PKDL) and guide the development of this project, I conducted an in-depth review of the most recent and relevant scientific literature. This review was fundamental in two ways: first, it provided essential clinical and immunological context to better interpret the structure and content of the dataset; second, it served as a reference point to validate and compare the analytical findings obtained during the project.

The following three articles were especially influential:

- **Younis et al. (2023)** [19]: This Phase II, open-label, randomized trial compared the safety and efficacy of PM/MF and LAmB/MF combination therapies in Sudanese PKDL patients. The study highlighted the impact of treatment duration and drug bioavailability on relapse and treatment outcomes. Its clinical findings supported the inclusion of treatment arms and liver function variables (e.g., ALT, bilirubin) in my analysis and offered a structured foundation for comparing treatment strategies within the dataset.
- **Torres et al. (2025)** [20]: This study applied unsupervised machine learning techniques to identify biomarkers associated with disease progression in PKDL. It emphasized the importance of variables such as 'Albumin g/L (Low:30;Upper:50)', 'Haematocrit %', and 'IFN-γ ' in distinguishing patient trajectories. It directly influenced the selection of key biomarkers and motivated the use of clustering and classification models in my own analysis.
- **Torres et al. (2024)** [6]: Focused on the cellular immune response during and after treatment, this article showed that low baseline levels of pro-inflammatory cytokines such as 'IFN-γ ', 'TNF-a', and 'IL-1B' were associated with relapse. These findings helped prioritize the inclusion of immunological variables in modeling and reinforced the idea that pre-treatment immune profiles can be predictive of long-term outcomes.

These works have not only shaped the methodology of this project but will also serve as reference points for discussion and interpretation of results.

A more detailed comparison between the findings of this analysis and the conclusions drawn in these studies is presented in later sections.

2.3. Dataset Description

The dataset contains clinical and laboratory information for 110 patients diagnosed with Post-Kala-Azar Dermal Leishmaniasis (PKDL), monitored across various time points up to 24 months. Each patient's data includes demographic variables, disease characterization, treatment information, lab test results, and immunological markers. The objective of analyzing this dataset was to identify patterns potentially associated with treatment response and relapse.

Given the complexity and clinical depth of the dataset, the variables were classified according to their type (categorical, numerical, temporal), and their clinical or biological relevance. Additionally, a null value analysis was conducted to determine data quality and guide future imputation strategies.

2.3.1. Variable Description and Clinical Relevance

In this section, we present a detailed description of the variables included in the dataset. The dataset includes a broad range of clinical, demographic, and laboratory variables. Each variable is described according to its name, type, possible values or ranges, and its clinical significance in the context of Post-Kala-Azar Dermal Leishmaniasis (PKDL).

2.3.2. Variable Description and Clinical Relevance

Below is a structured list of all variables included in the dataset, along with their data type, possible values or ranges, and clinical importance in the context of Post-Kala-Azar Dermal Leishmaniasis (PKDL).

2.3.2.1. Identifiers and Demographics

- **Sample_ID** (Categorical): Unique identifier for each sample, constructed by concatenating **Subject ID** and **Visit Day**. For example, for Subject **701** on Visit Day **000**, the corresponding *Sample_ID* is **701000**. This variable allows distinguishing between longitudinal samples from the same patient.

- **Subject ID** (Numerical): Unique identifier for each patient enrolled in the study. There are 110 patients in total with the following IDs: `701`, `702`, `703`, `...`, `810`. These IDs are used to track individual patient data across different visits.
- **Randomization arm** (Categorical): Assigned treatment group. Values: *ARM 1 PM (14 days) & MF (42 days)*, *ARM 2 AMBI & MF (28 days)*.
- **Gender** (Categorical): Biological sex. Values: *Male, Female*.
- **Age** (Numerical): Age in years. Range: 6–30.

2.3.2.2. Clinical Variables

- **PKDL Type/Characterisation** (Categorical): Lesion type. Values: *Macular, Papular, Plaque-like, Maculopapular*.
- **State of the disease** (Categorical): Indicates the patient's perception of disease evolution before recruitment. Values: *Worsening, Stable*. *Worsening* means the patient reported a progressive increase in lesion severity or distribution during the months prior to study inclusion.
- **Number of months with the disease** (Numerical): Duration of PKDL before treatment. Range: 4–152.
- **Did the patient complete the study** (Categorical): Whether the patient completed follow-up. Values: *Yes, No*.
- **Treatment outcome** (Categorical): Final result. Values: *Cured, Worsening, Improving, 0*.
- **Grade** (Ordinal): Skin lesion severity. Values: 0, 1, 2, 3.
- **VL_Diag** (Binary): Indicates whether the patient had a confirmed history of **VL (Visceral Leishmaniasis)** before developing PKDL. VL is a severe systemic parasitic disease caused by *Leishmania donovani*, characterized by fever, weight loss, hepatosplenomegaly, and pancytopenia. It is treated but often followed by dermal complications such as PKDL. Values: *1, N/A*.
- **VL_dur** (Numerical): Duration of VL in days. Range: 0–30.
- **VL_Drug** (Categorical): VL treatment received. Values: *SSG & PM, SSG alone, Unknown*.
- **Dist_Grade / Dens_Grade** (Ordinal): Clinical lesion distribution/density. Values: 0–3.

- **Type** (Categorical): Duplicate of lesion type. Same values as PKDL Characterisation.
- **Microscopy** (Categorical): Lab confirmation. Values: *Positive, Negative, N/A*.
- **Visit / Visit Day** (Categorical): Study follow-up point. Values: *Day 0, 42, 90, 180, 365*.
- **Study visit** (Categorical): Represents the scheduled or unscheduled time point at which each sample was collected during the study follow-up. This variable helps track the clinical and immunological evolution of patients over time. Possible values include:
  - *Screening* – Initial evaluation before treatment start.
  - *Day 1*, *Day 3*, *Day 7*, *Day 14*, *Day 28*, *Day 42* – Early treatment and short-term follow-up visits.
  - *Day 90*, *Day 180*, *Day 365* – Medium and long-term follow-up milestones.
  - *Unscheduled* – Additional or emergency visits not aligned with predefined protocol days.
- **Date sample taken** (Date): Indicates the exact calendar date when the biological sample was collected from the patient. Dates are recorded in the format `DD/MM/YYYY`, and span the years **2009**, **2018**, **2019**, **2020**, **2021**, and **2022**. This variable is essential for time-aligned analyses, longitudinal modeling, and tracking protocol adherence or delays in follow-up.

2.3.2.3. Biochemistry and Organ Function Markers

- **Albumin (g/L)** (Numerical): Serum albumin concentration, an important marker of nutritional and hepatic status. Observed range: **0 to 59.7 g/L**. Clinical reference range is typically between **30–50 g/L**. Low values may indicate malnutrition, liver dysfunction, or systemic inflammation, all of which are relevant in PKDL patient assessment.
- **Creatinine (mg/dL)** (Numerical): Serum creatinine level, a standard indicator of renal function. Observed range in the dataset: **0 to 1.8 mg/dL**. Clinical reference range: **0.2–1.4 mg/dL**. Elevated levels may indicate impaired kidney function, while low values can occur in states of reduced muscle mass or dilutional effects.

- **Potassium (mmol/L)** (Numerical): Serum potassium level, a key electrolyte involved in neuromuscular and cardiac function. Observed range in the dataset: **0 to 5.5 mmol/L**. Clinical reference range: **3–5 mmol/L**. Hypokalemia (low potassium) and hyperkalemia (high potassium) can both have serious implications, particularly in patients undergoing antileishmanial therapy.
- **SGOT/AST (U/L)** (Numerical): Serum aspartate aminotransferase level, an enzyme released in liver and muscle injury. Observed range: **0 to 436 U/L**. Clinical upper limit: $\leq$ **40 U/L**. Elevated AST levels can signal hepatocellular damage, which is crucial to monitor during potentially hepatotoxic treatments.
- **SGPT/ALT (U/L)** (Numerical): Serum alanine aminotransferase level, a liver-specific enzyme indicating hepatocellular injury. Observed range: **0 to 313 U/L**. Clinical upper limit: $\leq$ **40 U/L**. Like AST, elevated ALT can reflect liver stress or damage due to disease progression or drug side effects.
- **Total Bilirubin (mg/dL)** (Numerical): Serum bilirubin level, used to assess liver function and hemolytic activity. Observed range: **0 to 1.5 mg/dL**. Clinical upper limit: $\leq$ **1.2 mg/dL**. Elevated levels may suggest hepatic dysfunction, biliary obstruction, or increased red blood cell breakdown — all relevant for treatment monitoring in PKDL.

2.3.2.4. Grading and Clinical Significance (Biochemistry)

- **Albumin_grade** (Ordinal): Categorical grading of albumin levels. Observed values: **0, 2**. Used to stratify hypoalbuminemia severity; likely, *0 = normal*, *2 = abnormal*.
- **Creatinine_grade** (Ordinal): Grading of serum creatinine concentration. Observed values: **0, 1, 2**. May represent: *0 = normal*, *1 = mild elevation*, *2 = moderate/severe*.
- **Pottasium_grade** (Ordinal): Grading of potassium levels. Observed values: **0, 1, 3, 4**. Potentially reflects deviations from normal in both hypo- and hyperkalemia ranges.
- **SGOT/AST_grade** (Ordinal): Grading of AST enzyme levels. Observed values: **0, 1, 3**. *0 = within reference*, higher grades may reflect increasingly elevated liver injury markers.
- **SGPT/ALT_grade** (Ordinal): Grading of ALT enzyme levels. Observed

values: **0, 1, 2, 3**. Captures severity of hepatocellular damage; higher grades suggest greater deviation from normal.

- **Total Bilirubin_grade** (Ordinal): Grading of total bilirubin levels. Observed values: **0, 1**. *0 = within range*, *1 = elevated*.
- **Albumin_ClinSig** (Categorical): Clinical interpretation of albumin levels. Only value observed: **0**, likely indicating no clinically significant abnormality recorded.
- **Creatinine_ClinSig** (Categorical): Clinical significance of creatinine levels. Observed values: **0, Normal, Abnormal CS, Abnormal NCS**. Where:
  - *Normal* – within acceptable clinical limits.
  - *Abnormal CS* – Clinically significant abnormality.
  - *Abnormal NCS* – Abnormal, but not clinically significant.
  - *0* – Possible placeholder or default.
- **Pottasium_ClinSig** (Categorical): Clinical interpretation of potassium levels. Observed values: **0, Normal, Abnormal CS, Abnormal NCS**. Same scale as above.
- **SGOT/AST_ClinSig** (Categorical): Clinical significance of AST enzyme levels. Observed values: **0, Normal, Abnormal CS, Abnormal NCS**.
- **SGPT/ALT_ClinSig** (Categorical): Clinical significance of ALT enzyme levels. Observed values: **0, Normal, Abnormal CS, Abnormal NCS**.
- **Total Bilirubin_ClinSig** (Categorical): Clinical significance of total bilirubin levels. Observed values: **0, Normal, Abnormal NCS**. Notably, no cases of *Abnormal CS* were recorded for this marker.

2.3.2.5. Hematological Variables

- **Absolute Neutrophil (x10^3/µL)** (Numerical): Absolute neutrophil count, an indicator of innate immune activity. Observed range: **0 to 18.2 x10^3/µL**. Clinical reference range: **2.16–6.2 x10^3/µL**. Low counts (neutropenia) may reflect bone marrow suppression or disease-related immunosuppression, while high counts can indicate active infection or inflammation.
- **Basophils %** (Numerical): Percentage of basophils among total white blood cells. Observed values: **0, 1, 3 %**. Although typically a small frac-

tion of WBCs, elevated levels can be associated with allergic or parasitic responses.

- **Eosinophils %** (Numerical): Percentage of eosinophils in total leukocytes. Observed range: **0 to 10 %**. Eosinophilia is commonly associated with parasitic infections and allergic reactions, both potentially relevant in tropical diseases like PKDL.
- **Haematocrit %** (Numerical): Proportion of blood volume occupied by red blood cells. Observed range: **20 to 52 %**. Lower values indicate anemia, a common finding in chronic infectious diseases and a potential side effect of treatment.
- **Haemoglobin (g/dL)** (Numerical): Concentration of hemoglobin in blood, a key marker of oxygen-carrying capacity and anemia. Observed range: **9 to 16.3 g/dL**. Clinical reference range: **12–17.5 g/dL**. Values below reference suggest anemia, a frequent condition in patients with chronic parasitic infections or nutritional deficiencies.
- **Lymphocytes %** (Numerical): Proportion of lymphocytes in the white blood cell population. Observed range: **4 % to 80 %**. Lymphocytosis or lymphopenia may reflect immune response stages in PKDL or effects of immunomodulatory treatment.
- **Monocytes %** (Numerical): Percentage of monocytes in total leukocytes. Observed range: **0 % to 22 %**. Monocytes are involved in antigen presentation and chronic inflammation; elevations may occur in persistent infections.
- **Neutrophil %** (Numerical): Percentage of neutrophils in the white blood cell differential. Observed range: **13 % to 94 %**. High values are common in acute inflammation, while low percentages may suggest immune suppression.
- **Platelets x10^3/µL** (Numerical): Platelet count, essential for coagulation and immune response. Observed range: **43 to 881 x10^3/µL**. Clinical reference range: **150–500**. Thrombocytopenia may indicate bone marrow suppression or infection; thrombocytosis may appear in inflammation or recovery stages.
- **RBC x10^6/µL** (Numerical): Red blood cell count. Observed range: **3 to 7.1 x10^6/µL**. Alterations reflect anemia or polycythemia, both relevant in systemic disease context.
- **White Blood Cells x10^6/µL** (Numerical): Total white blood cell count.

Observed range: **4 to 37.9 x10^6/µL**. Clinical reference range: **4–10**. Leukocytosis may signal infection/inflammation; leukopenia can occur in immunosuppressed states.

2.3.2.6. Grading and Clinical Significance (Hematology)

- **Absolute Neutrophil_grade** (Ordinal): Observed values: **0 or .**. May reflect normality or missing/ungraded cases.
- **Basophils_%grade**, **Eosinophils_%grade**, **Haematocrit_%grade**, **Lymphocytes_%grade**, **Monocytes_%grade** (Ordinal): All observed values: **.** These variables appear unfilled or undefined in the current dataset.
- **Haemoglobin_%grade** (Ordinal): Observed values: **0, 1, 2, 3**. Reflects degree of anemia or deviation from normal.
- **Neutrophil_%grade** (Ordinal): Observed values: **0, 1, 2, 3, 4**. Captures neutrophil levels severity relative to clinical thresholds.
- **Platelets_grade** (Ordinal): Observed values: **0, 1, 2, 3**. Stratifies platelet count abnormalities.
- **RBC x10^6/µL_grade** (Ordinal): Observed values: **0, 1, 2, .**. Includes some undefined/missing cases.
- **White Blood Cells x10^6/µL_grade** (Ordinal): Observed values: **0, 1, 2**. Stratification of leukocyte abnormalities.
- **Absolute Neutrophil_ClinSig**, **Basophils%_ClinSig**, **Eosinophils%_ClinSig**, **Haematocrit%_ClinSig**, **Lymphocytes%_ClinSig**, **Monocytes%_ClinSig** (Categorical): All observed values: **0**. These variables appear to have no recorded clinical abnormalities in this dataset.
- **Haemoglobin_ClinSig**, **Neutrophil%_ClinSig**, **Platelets_ClinSig**, **RBC x106/µL_ClinSig**, **White Blood Cells x106/µL_ClinSig** (Categorical): Observed values: **0**, **Abnormal NCS**, **Abnormal CS**, **Normal**. These variables reflect expert clinical interpretation of the corresponding lab values.
  - **Normal**: Within expected clinical range.
  - **Abnormal CS**: Abnormal and Clinically Significant.
  - **Abnormal NCS**: Abnormal but Not Clinically Significant.
  - **0**: Likely a placeholder or non-evaluated entry.

2.3.2.7. Immunological and Inflammatory Markers

- **Granzyme B** (Numerical): A cytotoxic molecule secreted by activated cytotoxic T lymphocytes and natural killer (NK) cells. This biomarker reflects cellular immune activation and cytolytic potential. Values in the dataset span a broad range, including very low and markedly elevated concentrations, which may indicate inter-individual variability in immune response dynamics during or after treatment.
- **IFN-γ** (Numerical): Interferon-gamma, a key Th1 cytokine involved in macrophage activation and parasite clearance. High levels typically indicate strong cell-mediated immunity.
- **IL-1B** (Numerical): Interleukin-1 beta, a pro-inflammatory cytokine linked to innate immune activation and fever response.
- **IL-2** (Numerical): Interleukin-2, promotes T cell proliferation and differentiation, indicating adaptive immune activation.
- **IL-4** (Numerical): A Th2 cytokine involved in B cell activation and humoral immunity, often counterbalancing Th1 responses.
- **IL-5** (Numerical): Stimulates eosinophil activation and is associated with allergic and parasitic responses.
- **IL-10** (Numerical): An anti-inflammatory cytokine that downregulates pro-inflammatory signals; elevated levels may reflect immune regulation or suppression.
- **IL-13** (Numerical): Th2-related cytokine involved in tissue remodeling and allergic-type responses.
- **IL-17A** (Numerical): Produced by Th17 cells, plays a role in neutrophil recruitment and chronic inflammation.
- **IP10 (CXCL10)** (Numerical): A chemokine induced by IFN-γ, attracts activated T cells to sites of inflammation.
- **PDL1** (Numerical): Programmed death-ligand 1, an immune checkpoint protein that may reflect T-cell exhaustion or immune regulation during chronic infection.
- **TNF-α** (Numerical): A potent pro-inflammatory cytokine with roles in parasite control, but also associated with tissue damage when dysregulated.

- **IL-22** (Numerical): Contributes to epithelial barrier integrity and tissue repair; may be elevated in response to chronic skin inflammation.
- **IL-23** (Numerical): Supports maintenance of Th17 responses, contributing to chronic inflammatory processes.
- **TGF-β1** (Numerical): A regulatory cytokine with anti-inflammatory and fibrogenic roles; often associated with chronic disease progression or resolution.
- **Parasite load/µg DNA** (Numerical): Quantitative PCR measurement of *Leishmania* DNA in patient samples. Reflects direct parasitic burden and is used to monitor treatment response.
- **NLR (Neutrophil-Lymphocyte Ratio)** (Numerical): A systemic inflammation marker derived from hematological counts. Elevated values are associated with heightened inflammatory response or immune dysregulation.

2.3.3. Definition of new variables

At the beginning of the project, the dataset did not include explicit information on whether patients had relapsed or required rescue treatment. However, as the work progressed and additional clinical information became available, we identified a subset of patients for whom this outcome was known. Given the clinical relevance of these events and their potential utility for outcome prediction, we decided to define and integrate two new binary variables into the dataset: **Rescue** and **Relapse**.

The inclusion of these variables aimed to enhance the dataset's interpretability and enable classification approaches based on patient outcomes. This allowed us to stratify patients more meaningfully and to draw more solid and relevant conclusions about treatment effectiveness and disease progression.

- **Rescue**: Indicates whether a patient required rescue treatment during the study (value `1` if so, otherwise `0`). Rescue therapy was administered when the primary treatment failed or was not tolerated. A total of seven patients fell into this category. One patient (`Subject ID 702`) required rescue treatment due to hypersensitivity to liposomal amphotericin B, another (`Subject ID 713`) due to a lack of clinical response, and five others (`701, 729, 730, 752, 774`) as a result of relapse.
- **Relapse**: Indicates whether a patient experienced a relapse (value `1` if so, otherwise `0`). A relapse is defined as the recurrence of cutaneous

PKDL symptoms following an initial phase of clinical improvement after treatment. The five patients identified as having relapsed were the same ones who received rescue treatment due to disease recurrence: `701`, `729`, `730`, `752`, `774`.

These variables were created based on the corresponding Subject ID of each patient. If the Subject ID matched a known case of rescue or relapse, the respective variable was assigned a value of `1`; otherwise, it remained `0`. This approach allowed for integrating expert-derived outcome labels directly into the analysis pipeline.

2.4. Selection of Key Biomarkers

A subset of clinically and immunologically relevant biomarkers was selected from the dataset for focused analysis and predictive modeling. These variables, listed below, were chosen based on their biological relevance in the context of PKDL, their numerical nature (which facilitates statistical and machine learning techniques), and the availability of values across patients:

- **Cytokines and immune markers:** IFN-γ, IL-1B, IL-2, IL-4, IL-5, IL-10, IL-13, IL-17A, IL-22, IL-23, TGF-β1, IP10, PDL1, TNF-α.
- **Clinical biochemistry:** Albumin, Creatinine, Potassium, SGPT/ALT, Total Bilirubin.
- **Hematology:** Absolute Neutrophils, Basophils %, Eosinophils %, Haemoglobin, Lymphocytes %, Monocytes %, Platelets, RBC, White Blood Cells.
- **Ratios and direct parasitological assessment:** Neutrophil-to-Lymphocyte Ratio (NLR), Parasite load/μg DNA.

These biomarkers were prioritized due to their quantitative nature, which makes them directly suitable for statistical evaluation, correlation analysis, and model training.

In addition, these features represent diverse aspects of the disease process in PKDL: systemic and local inflammation, organ function, hematological response, and direct parasite burden.

In subsequent sections, we will describe how these biomarkers were used in exploratory analyses and incorporated into classification models to investigate their predictive potential with respect to relapse and treatment outcome.

Chapter 3

Methodology

This chapter summarizes the key steps of the analytical pipeline, including data preprocessing, exploratory analysis, clustering, supervised modeling, and statistical validation.

3.1. Data Preprocessing

This section describes the steps carried out to clean, structure, and prepare the dataset for analysis and modeling. The preprocessing pipeline was divided into three main stages: selecting relevant features, identifying or correcting outliers and handling missing values. These tasks were crucial in ensuring the consistency, interpretability, and analytical value of the dataset.

3.1.1. Feature Selection and Removal of Irrelevant Variables

Before addressing the issue of missing data, the first step in the preprocessing pipeline was to examine the structure and content of the dataset to identify and remove variables that were either redundant, non-informative, or lacked interpretability. This initial filtering step was essential to ensure that subsequent analyses, including imputation and modeling, would be performed on a clean and meaningful set of features.

Among the first variables removed was `Microscopy`, which contained largely missing or inconclusive values, and whose biological interpretation could not be clarified. Similarly, a set of grading and clinical significance variables

were discarded due to having only constant values such as `0`, `.`, or `N/A`, making them uninformative for analysis. These included:

- Grading fields: `Basophils %_grade`, `Eosinophils %_grade`, `Haematocrit %_grade`, `Lymphocytes %_grade`, `Monocytes %_grade`.
- Clinical significance fields: `Absolute Neutrophil_ClinSig`, `Basophils %_ClinSig`, `Eosinophils %_ClinSig`, `Haematocrit %_ClinSig`, `Lymphocytes %_ClinSig`, `Monocytes %_ClinSig`.

Additionally, the column `visit` was removed because it was a direct duplicate of `visit_day`, which provided the same information in a more complete and reliable format.

By removing these variables at the outset, the dataset was simplified and streamlined, reducing noise and improving the interpretability and efficiency of the following data preprocessing steps.

3.1.2. Outlier Detection and Treatment

Outlier detection was not carried out with the intention of eliminating these extreme values from the dataset, but rather to better understand the distribution of key numerical variables prior to handling missing data. Given the clinical nature of the dataset, many of these outliers may reflect true physiological variability or clinically relevant extremes. Therefore, no outlier removal was applied, in order to preserve the integrity and heterogeneity of the data.

Instead, the detection of outliers served as a critical exploratory step to guide the imputation strategy applied in the subsequent phase—particularly the method based on typical values such as medians and modes. Since many variables exhibited skewed distributions or extreme values, central tendency measures like the `median` were preferred over the mean for numerical imputation. This choice minimized the influence of outliers on imputed values, ensuring a more robust and unbiased representation of the underlying data.

Outlier detection was especially relevant for this imputation pathway, as many numerical variables exhibited skewed distributions or extreme values. In such cases, relying on the `mean` could introduce bias, pulling imputed values toward outliers and distorting the central tendency. Instead, the use of the `median` was favored for its robustness to extreme observations, resulting in a more stable and representative handling of missing values.

By understanding where outliers exist and how they influence the data distribution, we ensured that our imputation methods preserved the integrity and interpretability of the dataset.

Next, we present a summary of the outlier detection results across the key numerical variables included in the analysis. Table 3.1 shows the total number of outliers identified in each variable, highlighting those features where extreme values are more prevalent.

Variable	Number of Outliers
age	76
number of months with the disease	7
grade	8
VL_dur	136
dist_Grade	6
dens_Grade	6
Albumin g/L (Low:30;Upper:50)	109
Creatinine mg/dL (Low:0.2;Upper:1.4)	12
Pottasium mmol/L (Low:3;Upper:5)	227
SGOT/AST U/L (Upper:<=40)	17
SGPT/ALT U/L (Upper:<=40)	26
Total Bilirubin mg/dL (Upper:<=1.2)	13
Granzyme B	35
IFN-γ	32
IL-1B	26
IL-2	30
IL-4	22
IL-10	31
IL-13	27
IL-17A	36
TNF-a	41
Parasite load/µg DNA	95
NLR	53

Table 3.1: Summary of outliers detected per variable

Given the widespread presence of outliers across many numerical variables, we opted to impute missing values using the **median**, rather than the mean. The median is more robust in the presence of extreme values and ensures that the central tendency used in imputation is not distorted by skewed distributions. This decision was guided by the following rationale:

- Use the **mean** when data are symmetrically distributed and free from extreme values.
- Use the **median** when data are skewed or contain significant outliers, which was the case for most variables in this dataset.

Numerical variables with missing values and outliers:

```
['age', 'number of months with the disease', 'grade', 'VL_dur',
'dist_Grade', 'dens_Grade', 'Albumin g/L (Low:30;Upper:50)',
'Creatinine mg/dL (Low:0.2;Upper:1.4)', 'Pottasium mmol/L
(Low:3;Upper:5)', 'SGOT/AST U/L (Upper:<=40)', 'SGPT/ALT U/L
(Upper:<=40)', 'Total Bilirubin mg/dL (Upper:<=1.2)',
'Albumin_grade', 'Creatinine_grade', 'Pottasium_grade',
'SGOT/AST_grade', 'SGPT/ALT_grade', 'Total Bilirubin_grade',
 'Absolute Neutrophil (LowL:2.16;UpperL:6.2)', 'Basophils%',
 'Eosinophils%', 'Haematocrit%', 'Haemoglobin g/dL (LowL:12
 ;UpperL:17.5)', 'Lymphocytes%', 'Monocytes%', 'Neutrophil%',
 'Platelets x103/µL (LowL:150;UpperL:500)', 'RBC x106/µL',
 'White Blood Cells x106/µL (LowL:4;UpperL:10)',
 'Haemoglobin_grade', 'Neutrophil%_grade', 'Platelets_grade',
 'White Blood Cells x106/µL_grade', 'Granzyme B', 'IFN-γ',
 'IL-1B', 'IL-2', 'IL-4', 'IL-5', 'IL-10', 'IL-13', 'IL-17A',
 'IP10',  'PDL1', 'TNF-a',  'IL-22', 'IL-23', 'TGF-B1',
 'Parasite load/µg DNA', 'NLR']
```

Numerical variables with missing values and no outliers:

```
['VL_Diag', 'Albumin_ClinSig']
```

3.1.3. Handling Missing Values

3.1.3.1. Time-Based Missingness and Study Protocol

Before addressing missing values in the dataset, it is important to consider the context in which these variables were recorded. The clinical trial protocol defines specific time points at which each variable was collected. Figure 3.1 summarizes this schedule, based on the “Schedule of Assessments” outlined in the study documentation.

Table 3 Schedule of assessments

Protocol activities	Screening	Treatment period					EOT	Follow-up period		
	D-30 to D0	D1	D3	D7	D14	D28	D42	3M	6M	12M
Consent form	X									
Inclusion and exclusion criteria	X									
Demographic data and medical history	X									
Vital signs and physical exam	X	X	X	X	X	X	X	X	X	X
Audiometric test	X						X[a]		X[a]	X[a,b]
HIV test	X									
Pregnancy test [c]	X						X	X	X	X[a]
Depo-Provera® injection [c]	X							X	X[a]	
Hematology Hb, RBC, WBC, platelets	X		X	X	X	X	X	X[b]	X	X
Chemistry Albumin, ALT, AST, bilirubin, creatinine, potassium*	X	X	X	X	X	X	X	X[b]	X[b]	X[b]
Skin slit smear for microscopy and qPCR	X[d]						X	X	X	X
Tape Disc	X						X			
Blood PD (qPCR)	X						X	X	X	X
Skin biopsy for PK				X[e]	X[e]	X[e]	X[e]			
Blood sample for PK										
PK amphotericin B		X[f]		X[g]						
PK PM		X[f]			X[g]					
PK MF		X		X	X	X	X	X		
Blood sample for immunological parameters	X						X		X	
*PKDL evolution assessment, including photographs***	X						X	X	X	X
Safety assessment	SAEs	------- SAEs and AEs monitoring -------						S/AEs[h]		
Study treatment		---------- AmB + MF ----------								
		---------- PM + MF ----------								

AE = Adverse event; ALT = Alanine aminotransferase; AmB = AmBisome®; AST = Aspartate aminotransferase; D = Day; EOT = End of treatment; Hb = Hemoglobin; HIV = Human immunodeficiency virus; M = Month; MF = Miltefosine; PCR = Polymerase chain reaction; PD = Pharmacodynamic(s): PK = Pharmacokinetic(s); PKDL = Post-kala-azar dermal leishmaniasis; PM = Paromomycin;

Figure 3.1: Correlation Matrix: Model-Based Imputation and Typical Value Imputation

From this schedule, we can deduce the following:

- **PKDL clinical evaluation variables** such as `Grade`, `Dist_Grade`, `Dens_Grade`, and `Type` were only assessed at specific visits: Day 0 (pre-treatment), Day 42 (end of treatment), and during long-term follow-up (Days 90, 180, and 365).
- **Historical variables** like `VL_Diag`, `VL_Dur`, and `VL_Drug` were collected only at baseline (Screening or Day 0), as they relate to prior clinical history.

This explains why certain variables appear to have missing values at many time points: these values are not truly "missing" in the conventional sense, but rather were **not expected to be collected** according to the study design.

Moreover, a clear pattern was observed: samples missing `Grade` also lacked values for `Dist_Grade`, `Dens_Grade`, `Type`, `VL_Diag`, `VL_Dur`, and `VL_Drug`. This reinforces the idea that these features are jointly absent in certain visits, aligning with the time-point structure defined in the protocol.

Implication for imputation: Given that these variables are measured only at fixed time points, we should not treat their absence as random missingness. Instead, imputation should be limited to plausible clinical timepoints (e.g., Day 0 and Day 42), and based on contextually appropriate strategies such as mode imputation for categorical variables or domain-informed estimation for ordinal scores.

Therefore, in the preprocessing pipeline, we applied a unified strategy to this subset of variables. For time points where their presence was not expected (e.g., Day 3, Day 7), no imputation was performed. For cases where data was expected but absent (e.g., missed visit), we used the mode or median depending on variable type and observed distribution.

This protocol-aware handling ensures that imputations do not introduce artificial trends or biases misaligned with the clinical study design.

As a first step in the preprocessing pipeline, a thorough inspection of missing data was carried out to assess its extent and distribution across the dataset. Table 3.2 provides a detailed summary of all variables with missing values, including their data type, the absolute number of missing entries, and the percentage they represent relative to the total number of observations.

This preliminary analysis was crucial to inform subsequent decisions on how to handle missing data appropriately depending on the variable type, relevance, and degree of incompleteness. Based on this overview, specific

strategies were applied, as described below.

Variable	Type	Missing Count	Missing %
state of the disease	Categorical	928	89.40
number of months with the disease	Numerical	928	89.40
PKDL Type/Characterisation	Categorical	928	89.40
Microscopy	Categorical	928	89.40
did the patient complete the study	Identifier	830	79.96
treatment outcome	Categorical	830	79.96
IL-13	Numerical	744	71.68
IL-5	Numerical	744	71.68
IP10	Numerical	744	71.68
IL-17A	Numerical	744	71.68
IL-10	Numerical	744	71.68
IL-1B	Numerical	744	71.68
IL-4	Numerical	744	71.68
IL-2	Numerical	744	71.68
TNF-a	Numerical	744	71.68
IFN-γ	Numerical	744	71.68
Granzyme B	Numerical	744	71.68
PDL1	Numerical	744	71.68
IL-22	Numerical	744	71.68
TGF-B1	Numerical	744	71.68
IL-23	Numerical	744	71.68
Parasite load/µg DNA	Numerical	563	54.24
VL_Drug	Categorical	553	53.28
grade	Grading	553	53.28
visit	Identifier	553	53.28
type	Categorical	553	53.28
dens_Grade	Grading	553	53.28
dist_Grade	Grading	553	53.28
VL_dur	Numerical	553	53.28
VL_Diag	Numerical	553	53.28

Table 3.2: Summary of variables with missing data and data types

Initially, missing values were handled based on domain knowledge and logical inference. For instance, missing entries in the columns `visit` and `visit_day` were completed by extracting the last characters from the `Sample_ID`, which encode visit information. When `Sample_ID` contained entries like `708Unscheduled`, these were interpreted as `visit = Unscheduled`. Since `visit` and `visit_day` were redundant, only the `visit_day` column was retained.

Missing `Subject ID` values were inferred from the first three digits of the corresponding `Sample_ID`. Additionally, several columns with only a single

constant value or systematically missing entries (e.g., `Microscopy`, `Basophils %_grade`, `Eosinophils %_grade`,`Absolute Neutrophil_ClinSig`) were removed due to their lack of variability or clinical interpretability.

Some clinical significance fields, such as `Haemoglobin_ClinSig`, `Platelets_ClinSig`, and `Neutrophil %_ClinSig`, were filled with the value "`not measured`" to standardize missing information across the dataset.

Two distinct imputation strategies were then applied and compared:

1. **Model-based imputation:** This method involved using machine learning models to predict and impute missing values, leveraging the rest of the dataset as input features [21]. For each variable with missing values, a separate model was trained, based on whether the target variable was numerical or categorical:

   - If the variable was **numerical** (e.g., *Creatinine, Albumin, Granzyme B*), a *Random Forest Regressor* was used.
   - If the variable was **categorical or ordinal** (e.g., *treatment outcome, albumin_grade*), or had a small number of unique values, a *Random Forest Classifier* was applied.

   To avoid overfitting and enhance model robustness, each model was trained on the subset of complete cases (rows where the target variable was observed). Columns with excessive missingness (more than 90 %) were excluded from the modeling process. In our case, as shown in Table 3.2, none of the variables exceeded this missingness threshold. Additionally, irrelevant identifiers such as `Sample_ID`, `Subject ID`, and `Date sample taken` were removed before training.

   For preprocessing, numerical predictors were imputed with the mean, while categorical variables were imputed with the most frequent category and encoded using one-hot encoding. Each pipeline included both preprocessing and modeling steps, ensuring consistency.

   The trained pipeline then predicted the missing values for the respective variable, which were subsequently integrated back into the dataset. This approach allowed for context-aware imputations that leveraged complex interactions among variables.

2. **Typical value imputation:** As an alternative strategy and for comparison purposes, we also implemented a more traditional imputation method using typical values [22]. For **numerical variables**, missing values were filled with the **median**, since many of these variables we-

re found to contain outliers and exhibited skewed distributions. This decision was made to reduce the influence of extreme values on the imputation process.

For **categorical variables**, the **mode** (most frequent category) was used as a representative value. This approach assumes that the data was *missing at random (MAR)*, and follows the rationale discussed in the supporting literature, where missing entries were often attributed to practical limitations (e.g., laboratory capacity, time constraints).

The two resulting datasets were saved separately and will be used independently in subsequent analysis for comparison.

3.2. Exploratory Data Analysis

3.2.1. Selective Undersampling Strategy (SUS)

In order to address the significant class imbalance present in the dataset—specifically, the limited number of relapse cases compared to cured patients—we applied a selective undersampling strategy (SUS). Unlike random undersampling, which removes instances arbitrarily, SUS retains both the most representative and the most atypical (outlier) samples from the majority class. This allows the dataset to be balanced without sacrificing clinical diversity or informative edge cases [23].

Why SUS? This method was particularly suitable for our study because:

- The dataset is relatively small, and random removal of cured patients risks discarding clinically valuable information.
- Many features are biomedical or immunological markers whose variability is relevant for prediction and interpretation.
- Maintaining a representative distribution of the cured class is critical, especially in a scientific context where model interpretability and biological plausibility are as important as predictive performance.

How was it applied? We first separated the dataset into relapse cases (`Relapse` = 1) and cured cases (`Relapse` = 0). For the cured group, we computed the local density of each sample using the mean distance to its 5 nearest neighbors, based on a set of biologically relevant features selected in Section 2.4 These included hematological and biochemical markers such as *SGPT/ALT*, *Haemoglobin*, *Neutrophils %*, *IL-10*, *TNF-a*, *NLR*, and others.

The samples with the highest density were considered *representative cured cases*, while those with the lowest density were treated as *edge-case cured patients* (potential outliers). To construct the balanced dataset, we selected:

- All relapse cases ($n = 48$)
- The top 24 densest cured cases
- The bottom 24 sparsest cured cases

The resulting balanced dataset (`df_SUS`) contained 96 samples and preserved the internal variability of the original data while correcting the severe class imbalance.

It is important to note that undersampling was not used to impute missing values. Instead, its role was strictly for class balancing prior to applying machine learning models and clustering techniques. By reducing redundant majority samples and keeping clinically informative cured cases, SUS improved the fairness and robustness of downstream analyses.

Further reading and justification for this technique can be found in the work of Fernández et al. [23], who propose density-aware undersampling as a powerful method in imbalanced regression and classification tasks.

3.2.2. Correlation Analysis

To assess linear relationships between variables in the dataset, we computed the Pearson correlation coefficient. This method evaluates the degree of linear association between pairs of continuous variables, returning a value between -1 and 1. Values close to 1 indicate strong positive linear correlation, values near -1 indicate strong negative correlation, and values around 0 suggest no linear relationship [24]

The Pearson correlation was chosen because:

- Most of the variables analyzed are numerical or ordinal in nature, which aligns well with Pearson's assumptions.
- The method is widely used in exploratory data analysis and provides a clear, interpretable metric for identifying redundant features.
- It is particularly effective at detecting multicollinearity, which can negatively affect clustering or dimensionality reduction algorithms.

A full pairwise correlation matrix was computed, and all variable pairs with an absolute correlation coefficient greater than 0.8 were flagged for review.

To avoid redundancy and simplify the space of features, these highly correlated pairs were carefully evaluated and, in selected cases, one of the variables was removed. This process is detailed in the next section.

3.2.2.1. Correlation Matrix: Model-Based Imputation and Typical Value Imputation

To simplify the presentation and avoid redundancy, we grouped the correlation results from both imputed datasets (one using machine learning-based imputation and the other with typical values) into a single section, as they produced nearly identical outcomes. This unified analysis allows for streamlined interpretation without compromising the robustness of the insights.

Figure 3.2 shows the correlation matrix computed using the Pearson correlation coefficient. As expected, several variables exhibited strong pairwise correlations, indicating potential multicollinearity. Notable correlations include:

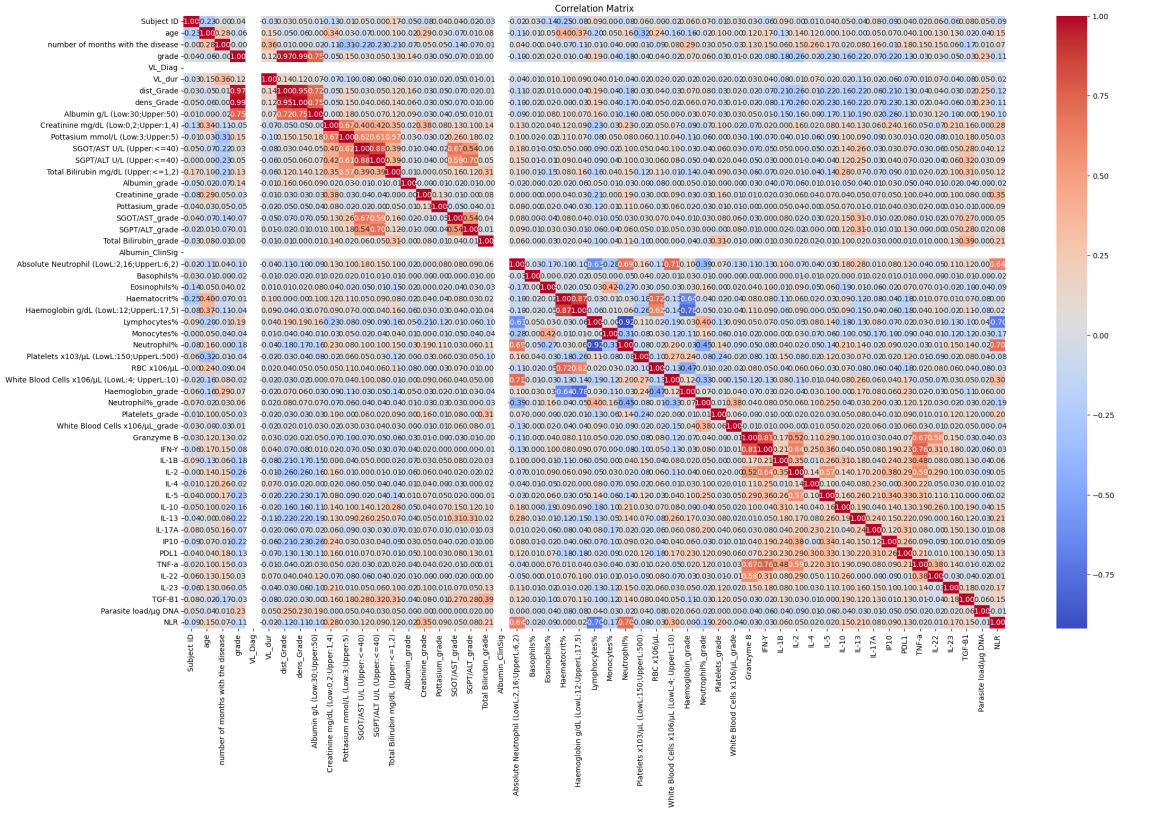


Figure 3.2: Correlation Matrix: Model-Based Imputation and Typical Value Imputation

- **Grade – dens_Grade:** $r = 0.99$
- **Grade – dist_Grade:** $r = 0.97$
- **dens_Grade – dist_Grade:** $r = 0.95$
- **SGOT/AST – SGPT/ALT:** $r = 0.88$
- **Haematocrit % – Haemoglobin:** $r = 0.87$
- **Lymphocytes % – Neutrophil %:** $r = -0.92$
- **Granzyme B – IFN-γ:** $r = 0.81$
- **Rescue – Relapse:** $r = 0.89$

Eliminating highly correlated features before applying unsupervised learning algorithms, such as clustering, is a widely recommended practice. This avoids redundancy, reduces dimensional distortion in the feature space, and improves the interpretability and stability of the resulting clusters.

Justification for Feature Removal:

1. **Haematocrit % vs Haemoglobin (r = 0.87)** Both measure related hematological parameters: hemoglobin reflects oxygen-carrying capacity, while hematocrit indicates red blood cell volume. Since hemoglobin is more commonly referenced in infectious diseases like PKDL as an anemia marker, we opted to retain it and remove **Haematocrit %**.
2. **Lymphocytes % vs Neutrophil % (r = –0.92)** These variables are inversely related as they are part of the leukocyte differential count. Given the existence of the derived variable NLR (neutrophil-to-lymphocyte ratio), and the immunological relevance of lymphocytes in chronic diseases like PKDL, we retained **Lymphocytes %** and dropped **Neutrophil %**.
3. **Grade, dens_Grade, dist_Grade** These three variables represent similar measures of lesion severity and distribution. Due to their extremely high correlations (up to 0.99), we retained **Grade** and removed the other two.
4. **SGOT/AST vs SGPT/ALT** Both are liver enzymes with overlapping clinical meaning. Since ALT (SGPT) is more liver-specific, we retained it and removed **SGOT/AST**.
5. **Granzyme B vs IFN-γ** While both are relevant immune markers, IFN-γ has more consistent associations with PKDL progression according to the literature. We thus removed **Granzyme B**.

The following variables were removed as a result of this correlation filtering step:

```
['Granzyme B', 'dist_Grade', 'dens_Grade', 'SGOT/AST U/L
(Upper:<=40)',  'Haematocrit%', 'Neutrophil%']
```

3.2.2.2. Correlation Matrix: Post-Undersampling Analysis

After applying an undersampling technique to balance the classes (specifically the `Recaida` variable), we repeated the correlation analysis on the reduced dataset. This allowed us to verify whether key relationships between features remained consistent and to identify any new significant correlations introduced by changes in the data distribution.

Figure 3.3 shows the Pearson correlation matrix computed after undersampling. The overall structure remained similar to that of the original dataset; however, some correlations became stronger or more pronounced, especially among immunological markers and derived variables.

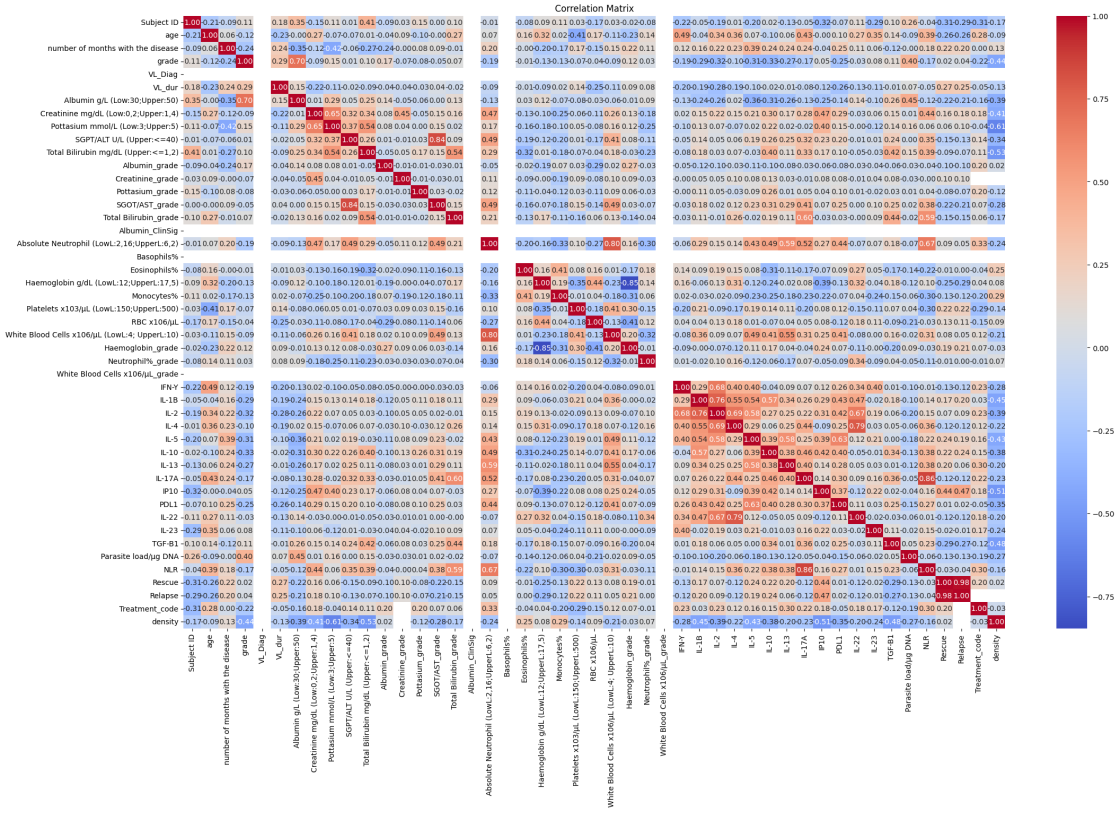


Figure 3.3: Correlation Matrix Computed After Undersampling

Several feature pairs exceeded the absolute correlation threshold of 0.8, suggesting redundancy or tightly linked biological processes. The most relevant correlations observed were:

- **Grade – dens_Grade:** $r = 0.98$
- **Grade – dist_Grade:** $r = 0.97$
- **SGOT/AST – SGPT/ALT:** $r = 0.97$
- **SGPT/ALT – SGPT/ALT_grade:** $r = 0.88$
- **SGOT/AST – SGPT/ALT_grade:** $r = 0.84$
- **SGPT/ALT – SGOT/AST_grade:** $r = 0.84$
- **SGOT/AST_grade – SGPT/ALT_grade:** $r = 0.87$
- **Absolute Neutrophil – Neutrophil %:** $r = 0.81$
- **Haematocrit % – Haemoglobin:** $r = 0.86$
- **Haemoglobin – Haemoglobin_grade:** $r = -0.85$
- **Lymphocytes % – Neutrophil %:** $r = -0.95$
- **Platelets_grade – IL-17A:** $r = 0.87$
- **Platelets_grade – NLR:** $r = 0.86$
- **Granzyme B – IFN-γ:** $r = 0.96$
- **Granzyme B – TNF-a:** $r = 0.86$
- **IFN-γ – TNF-a:** $r = 0.85$
- **IL-2 – TNF-a:** $r = 0.85$
- **IL-17A – NLR:** $r = 0.86$
- **Rescate – Recaida:** $r = 0.98$

As a general rule, variables with correlation coefficients above 0.90 or below –0.90 are considered candidates for removal, especially in models sensitive to multicollinearity (e.g., logistic regression, neural networks). Removing such features:

- Avoids redundancy and reduces dimensionality.
- Preserves model interpretability and statistical stability.
- Prevents biased clustering due to overrepresented information.

Variables removed due to redundancy and domain considerations:

- **Grade, dist_Grade, dens_Grade:** highly correlated ($r > 0.97$). Only `Grade` was retained as it was the original categorical severity indicator.
- **SGOT/AST, SGPT/ALT and their grading versions:** redundant values ($r > 0.84$). Raw lab values were discarded in favor of interpretable grade variables.
- **Haematocrit % vs. Haemoglobin:** $r = 0.86$. `Haematocrit %` was removed, as haemoglobin is more widely used in PKDL and similar inflammatory contexts.
- **Lymphocytes % vs. Neutrophil %:** strongly inversely correlated ($r = -0.95$). `Neutrophil %` was dropped since the neutrophil–lymphocyte ratio (NLR) was also included.
- **Granzyme B, IFN-γ, TNF-a:** highly correlated cytokines ($r > 0.85$). `Granzyme B` and `TNF-a` were removed to avoid redundancy.
- **NLR:** correlated with several immune features ($r > 0.85$). Removed to avoid overlap with its component variables.

Final set of removed variables:

`Granzyme B, dist_Grade, dens_Grade, SGOT/AST U/L (Upper:<=40), SGPT/ALT U/L (Upper:<=40), Haematocrit %, Neutrophil %, TNF-a, NLR`

This cleaning step ensured that the reduced dataset used for clustering and prediction would be free from strong multicollinearity, while still preserving clinical and biological relevance.

3.2.3. Variable Transformation

To ensure that all features in the dataset were compatible with downstream machine learning algorithms—both unsupervised and supervised, we performed a series of variable transformations. Many of the original features were categorical, ordinal, or stored as strings, which required conversion into numerical formats.

The main types of transformations applied were:

- **Ordinal cleaning and conversion:** Some columns, such as `RBC x106/µL_grade`, contained non-numeric placeholders (e.g., `"."`). These rows

were filtered out, and the column was cast into numeric type to be properly used in analysis.

- **Binary encoding:** Binary categorical variables including `Gender`, `state of the disease`, and `did the patient complete the study` were encoded numerically as 0 or 1 using `pandas`' built-in categorical tools.
- **Treatment arm coding:** The `Randomization arm` variable was mapped into a new numeric feature named `Tratamiento_cod`, where `ARM 1 PM & MF` was encoded as 0 and `ARM 2 AMBI & MF` as 1. This transformation was essential for distinguishing treatment groups in modeling tasks.
- **One-hot encoding:** For nominal variables with more than two categories (e.g., `PKDL Type/Characterisation`, `VL_Drug`, or clinical significance scores), one-hot encoding was applied using `pandas.get_dummies`. The first category of each feature was dropped to avoid multicollinearity.

These transformations ensured that all variables were in a numerical format suitable for modeling, and preserved categorical structure where relevant. This step was critical for enabling feature scaling, correlation analysis, and integration into clustering or supervised learning workflows.

3.2.4. Data Normalization and Scaling

In order to ensure that numerical features contributed equally to distance-based and gradient-based algorithms—such as KMeans clustering and logistic regression—we applied standardization to the dataset.

Standardization transforms features to have a mean of 0 and a standard deviation of 1. Unlike normalization (which rescales values into a fixed range such as [0, 1]), standardization is better suited to clinical datasets where features often span different units and scales, and where outliers or skewed distributions are present. This transformation is particularly beneficial for algorithms that assume Gaussian distributions or are sensitive to feature magnitude.

We opted to scale only the relevant clinical and biological variables, excluding identifiers (`Sample_ID`, `Subject ID`), dates (`Date sample taken`), and outcome labels (`Relapse`, `Rescue`), as well as purely categorical fields that had been encoded in previous preprocessing steps.

The standardization process was applied using the `StandardScaler` from

`scikit-learn`. The final transformed dataset preserved the structure of the original clinical variables while making them compatible with modeling techniques requiring scaled inputs. A sample of the resulting standardized values is shown in Table 3.3, confirming that all selected features were centered and scaled appropriately.

Preview of Standardized Numerical Features						
Index	**Age**	**Grade**	**Albumin**	**Creatinine**	**IL-2**	**TNF-a**
1	-0.38	-0.41	-0.34	0.29	-0.30	-0.22
5	-0.38	-0.41	-0.34	0.29	-0.29	0.14
6	-0.38	3.20	-0.34	0.61	-0.26	-0.75
7	-0.38	3.20	-0.34	-0.08	-0.29	-0.20
8	-0.38	3.20	-0.34	-1.30	-0.29	-0.74

Table 3.3: Example of standardized clinical features using `StandardScaler`. Values are centered around 0 with unit variance.

Note: Negative values (e.g., for age) result from the standardization process, which centers each variable around a mean of 0 and scales it by its standard deviation. Thus, a value like -0.38 *indicates that the observation lies 0.38 standard deviations below the dataset's average age.*

3.3. Unsupervised Learning

Unsupervised learning techniques were employed to uncover hidden patterns, intrinsic structures, and natural groupings within the dataset without the use of labeled outcomes. These methods are particularly useful in exploratory data analysis, anomaly detection, and pre-modeling stages where understanding data distribution and relationships is crucial [25] . This study primarily utilized dimensionality reduction and clustering methods to analyze the heterogeneity of the population and support subsequent modeling steps.

Dimensionality Reduction and Visualization

Principal Component Analysis (PCA) was first applied to assess the intrinsic dimensionality of the data and to inform the selection of components for downstream analysis. The *explained variance ratio* and *cumulative variance*

plots were examined to determine how much information was retained by each principal component [26]. To assist in this selection, we employed the elbow plot, a commonly used heuristic to identify the point at which adding further components yields diminishing returns in terms of explained variance.

Based on this evaluation:

- In both datasets (prior to undersampling), we selected the first **10 principal components**, which captured a substantial portion of the total variance while preserving interpretability.
- In the **undersampled dataset**, the elbow in the explained variance curve occurred much earlier. Therefore, we reduced the dimensionality to only **6 principal components**, which was sufficient to retain the most relevant structure of the data in this case.

All continuous variables were standardized using `StandardScaler` before applying PCA to prevent features with larger scales from dominating the analysis.

To further explore potential **non-linear relationships** within the data, we also employed two manifold learning techniques: **t-distributed Stochastic Neighbor Embedding (t-SNE)** and **Uniform Manifold Approximation and Projection (UMAP)**. These methods project high-dimensional data into two dimensions and are well-suited for visualizing clusters or patterns.

- **t-SNE** was used to capture local structure and visualize potential groupings between relapsed and cured patients.
- **UMAP** preserved both local and global structure, enabling a more continuous and topologically faithful representation of the sample distribution. Visualizations were colored by relapse status to facilitate interpretation.

Together, these dimensionality reduction approaches provided key insights into the structure of the data and supported subsequent unsupervised learning steps aimed at identifying **potential biomarkers of cure in PKDL**.

3.3.1. Clustering

Clustering techniques were utilized to identify subgroups within the population based on shared characteristics, without relying on prior labels. The goal was to discover natural partitions in the data that could inform later stages of supervised modeling or support clinical hypothesis generation [27].

A variety of clustering algorithms were evaluated to ensure robust results across different assumptions of cluster structure and density.

3.3.1.1. Clustering with K-Means

K-Means clustering was used as a baseline algorithm for identifying homogeneous groups in the dataset. It partitions the data into K clusters by minimizing the within-cluster sum of squared distances to the cluster centroids. The algorithm assumes spherical cluster shapes and requires pre-specifying the number of clusters [27].

To explore suitable values of K, we manually evaluated clusterings with values ranging from 2 to 6. For each K, K-Means was applied on the PCA-reduced feature space, and the resulting clusters were visualized using scatter plots of the first two principal components. This qualitative inspection helped assess how well-separated and interpretable the clusters appeared. Additionally, we computed the mean relapse rate within each cluster to identify potential clinical relevance and distinguishability among the groups. This combined visual and outcome-based approach guided our selection of an appropriate number of clusters. The best results were obtained with $K = 2$ on the undersampled dataset, where clusters were clearly separated and aligned with relapse behavior, and with $K = 5$ for the other two datasets, providing a balanced trade-off between separation and interpretability.

Despite its simplicity, K-Means is sensitive to initial centroid placement and is limited in its ability to identify non-convex clusters, which motivated the use of more flexible algorithms.

3.3.1.2. Clustering with DBSCAN and HDBSCAN

Density-based clustering algorithms such as DBSCAN and HDBSCAN were employed to capture more complex, non-spherical cluster shapes and to automatically identify noise points (outliers). These methods do not require specifying the number of clusters in advance and are particularly effective in handling datasets with varying cluster densities.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) groups together points that are closely packed and labels points in low-density regions as outliers. Its key parameters, `eps` (neighborhood radius) and `min_samples` (minimum points to form a dense region), were optimized through neighborhood distance plots and inspection of cluster results in 2D projections [28].

HDBSCAN (Hierarchical DBSCAN), an extension of DBSCAN, constructs a hierarchical clustering tree and extracts the most stable clusters, offering better performance on datasets with variable density. It was particularly valuable in our context, providing more interpretable and granular results with reduced sensitivity to parameter tuning [29].

Both algorithms were applied on the output of PCA and UMAP projections, leveraging lower-dimensional embeddings to improve computational efficiency and clustering accuracy. The `NearestNeighbors` algorithm was used to construct k-distance graphs and support DBSCAN parameter selection.

Cluster validity was assessed using internal metrics such as `silhouette_score`, and external metrics like `adjusted_rand_score` and `homogeneity_score` when applicable.

3.3.1.3. Cluster Interpretation and Relationship to Clinical Outcomes

Once stable clusters were identified, efforts were made to interpret them in terms of clinical and demographic variables. Descriptive statistics and visualizations (e.g., boxplots, heatmaps, violin plots) were used to characterize each cluster's profile across key features.

Categorical and numerical comparisons between clusters were conducted using appropriate statistical tests (e.g., `chi2_contingency`, `ttest_ind`, `mannwhitneyu`), allowing the identification of features that significantly differentiated the clusters.

Furthermore, clusters were linked to known clinical outcomes and risk factors where available, providing insights into whether the discovered subgroups corresponded to meaningful patterns such as treatment response, risk stratification, or disease progression.

The interpretability of clusters was enhanced using dimensionality reduction projections (PCA and UMAP), enabling intuitive visualization of cluster boundaries and overlaps. These analyses laid the groundwork for downstream predictive modeling, helping to refine feature selection and support model transparency.

3.4. Supervised Learning

Supervised learning is a machine learning approach that learns a function from labeled input-output pairs, allowing it to predict outcomes for new, unseen data. In our case, we framed the prediction of disease relapse as a

binary classification problem, where the target variable `Relapse` takes on values of 0 (no relapse) or 1 (relapse).

Given the clinical importance of accurately identifying patients at risk of relapse, we applied supervised learning techniques to model and analyze a wide set of clinical, biochemical, and immunological biomarkers. This approach enabled us to discover potentially relevant predictors and evaluate how well different models could distinguish between relapse and non-relapse cases.

Due to the significant class imbalance in our dataset ($0 = 990$, $1 = 48$), we employed various techniques—such as SMOTE (Synthetic Minority Oversampling Technique)—to improve model performance and avoid biased predictions towards the majority class.

3.4.1. Feature Selection: Key Biomarkers

The predictor variables used in our models included a comprehensive set of clinical, biochemical, and immunological features, along with a binary indicator for treatment assignment. Specifically, the feature set `X` was constructed as follows:

- Clinical and laboratory biomarkers: `age`, `number_of_months_with_the_disease`, `VL_Diag`, `Albumin_g_L__Low_30_Upper_50_`, `Creatinine_mg_dL__Low_0_2_Upper_1_4_`, `Pottasium_mmol_L__Low_3_Upper_5_`, `SGPT_ALT_U_L__Upper___40_`, `Total_Bilirubin_mg_dL__Upper___1_2_`, `Absolute_Neutrophil__LowL_2_16_UpperL_6_2_`, `Haemoglobin_g_dL__LowL_12_UpperL_17_5_`, `Platelets_x103_µL__LowL_150_UpperL_500_`, `RBC_x106_µL`, `White_Blood_Cells_x106_µL__LowL_4__UpperL_10_`, etc.
- Derived and categorized variables: `Albumin_grade`, `Creatinine_grade`, `Pottasium_grade`, `SGOT_AST_grade`, `SGPT_ALT_grade`, `Total_Bilirubin_grade`, `Albumin_ClinSig`, and other clinically meaningful categorizations.
- Cytokines and immunological markers: `IFN_Y`, `IL_1B`, `IL_2`, `IL_4`, `IL_5`, `IL_10`, `IL_13`, `IL_17A`, `IL_22`, `IL_23`, `IP10`, `PDL1`, `TNF_a`, `TGF_B1`, `Parasite_load_µg_DNA`, and `NLR`.
- Treatment indicator: `Treatment_code`, a binary variable derived from the original `Randomization_Arm`, encoding the two treatment arms.

These features were selected based on both clinical relevance and statistical

contribution to relapse prediction, ensuring that the models were trained on biologically interpretable and informative inputs.

3.4.2. Models Tested

We evaluated multiple supervised learning models to predict relapse. Each model was tuned and tested with and without data balancing using SMOTE.

3.4.2.0.1. Logistic Regression with L1 Regularization

- **Implementation:** `LogisticRegressionCV` from `sklearn`.
- **Hyperparameters:** 10-fold cross-validation, L1 penalty (for sparsity/-feature selection), `solver='saga'`, `max_iter=3000`, `scoring='roc_auc'`.
- **Evaluation:** ROC AUC, classification report. Feature importance was derived from model coefficients.

3.4.2.0.2. Random Forest

- **Implementation:** `RandomForestClassifier` from `sklearn`.
- **Hyperparameters:** `n_estimators=100-200`, `class_weight='balanced'` in some cases.
- **Evaluation:** ROC AUC, confusion matrix, classification report. Feature importance from Gini impurity-based scores.

3.4.2.0.3. XGBoost

- **Implementation:** `XGBClassifier` from `xgboost`.
- **Hyperparameters:** `n_estimators=500-1050`, `max_depth=3`, `learning-_rate=0.0001-0.1`, `scale_pos_weight` to handle class imbalance, `eval_metric='auc' or 'logloss'`.
- **Evaluation:** Same as above. Feature importances plotted and ranked.

3.4.2.0.4. LightGBM and CatBoost

- **Implementation:** `LGBMClassifier` and `CatBoostClassifier`.
- **Hyperparameters:** Tuned via standard grid settings (e.g., depth, leaves, regularization, learning rate).
- **Evaluation:** ROC AUC, F1 Score, SHAP values for model explainability.

3.4.2.0.5. Support Vector Machine (SVM)

- **Implementation:** `SVC` with RBF kernel.
- **Hyperparameters:** `C=1`, `class_weight='balanced'`, `gamma='scale'`.
- **Evaluation:** ROC AUC, classification report. Feature interpretation through SHAP.

3.4.3. Balanced vs. Unbalanced Dataset

The original dataset was highly imbalanced ($0 = 990$, $1 = 48$, where 1 indicates relapse). To address this, we experimented with the Synthetic Minority Oversampling Technique (SMOTE) in different scenarios:

- **SMOTE applied before train-test split:** produced a fully balanced dataset for both training and testing.
- **SMOTE applied only to training set:** test set remained naturally imbalanced, providing more realistic evaluation.
- **StratifiedKFold with SMOTE:** applied SMOTE before performing 5-fold stratified cross-validation to assess generalizability and avoid overfitting.

3.4.3.0.1. Metrics Used

- **Classification Report:** Includes precision, recall, F1-score, and support for each class.
- **ROC AUC Score:** Measures the ability of the model to distinguish between classes.
- **Confusion Matrix:** To visualize the distribution of predictions.
- **Feature Importance and SHAP Analysis:** For model interpretability and transparency.

We observed performance improvements when using SMOTE, particularly in recall for the minority class. However, balancing on the test set tended to inflate metrics, emphasizing the need for caution when interpreting results from fully resampled datasets.

3.5. Parametric and Non-Parametric Tests

To further validate the importance of the most relevant biomarkers identified by the supervised learning models, we conducted both parametric and non-parametric statistical tests. These were aimed at verifying whether the distributions of these variables significantly differed between patients who relapsed and those who did not (`Relapse` = 1 vs. `Relapse` = 0).

3.5.1. p-value

For each selected feature, we computed the p-value using appropriate statistical tests. A low p-value (typically < 0.05) indicates that the variable exhibits a statistically significant difference between the relapse and non-relapse groups. These p-values were later corrected for multiple comparisons using the False Discovery Rate (FDR) method to reduce the chance of Type I errors.

We also fitted a logistic regression model using `statsmodels` on the selected variables to estimate their individual contribution (odds ratios and confidence intervals) to the likelihood of relapse:

- Logistic regression with `Relapse` as the binary target.
- Coefficients were transformed into odds ratios using `exp(coef)`.
- Confidence intervals and p-values were obtained for each feature.

3.5.2. t-test

After identifying variables with low p-values, we further validated them using independent two-sample t-tests, assuming unequal variances (Welch's t-test). This test evaluates whether the mean values of a feature differ significantly between the relapse and non-relapse groups.

- Applied only to numeric variables with sufficient data in both groups.
- For each variable, we computed the t-statistic and its corresponding p-value.
- We corrected the resulting p-values using the Benjamini–Hochberg FDR procedure.

In addition, we visualized the most statistically significant variables (FDR-adjusted p-value < 0.05) using boxplots, comparing their distributions between the two outcome classes. This allowed us to interpret their discriminative power not only through numerical metrics but also visually.

Post-Analysis and Comparison

Following this statistical validation, we applied all previously described clustering algorithms exclusively to the set of variables found to be statistically significant. The goal was to explore whether these variables could also group patients in a meaningful and unsupervised manner, potentially revealing latent patterns related to relapse.

Finally, the patterns and biomarkers identified through our pipeline were compared with those reported in our reference publication, in order to evaluate the consistency and reproducibility of findings across studies. This comparative step aimed to determine whether our model-driven approach leads to conclusions that align with established clinical knowledge and prior research.

Exploratory Analysis of Non-Model Variables

In addition to the features selected by our supervised learning models, we also performed statistical analysis on a separate set of variables that were not highlighted by the models but were considered important for clinical interpretation or study design.

These included both categorical and numerical variables that may influence relapse or provide context to the patient's condition. The analysis was conducted on all patients for whom the target variable `Relapse` was defined.

- **Categorical Variables:** For each categorical variable, we performed a Chi-squared test of independence (χ^2) to assess whether its distribution differed significantly between relapse and non-relapse groups.
- **Numerical Variables:** For continuous features, we applied Welch's t-tests (assuming unequal variances) to compare the means between the two groups.

The results were sorted by p-value, and those variables with statistically significant differences ($\texttt{p-value} < 0.05$) were visualized using:

- Count plots for categorical variables, showing the distribution of categories by relapse status.
- Boxplots for numerical variables, comparing distributions across relapse outcomes.

This complementary analysis allowed us to include expert-driven variables in our assessment and determine whether their distributions supported a

potential role in relapse prediction. Although not directly selected by the predictive models, some of these variables may hold interpretative or causal value. Therefore, they were also considered in downstream visualizations and interpretation.

In summary, this exploratory layer helped ensure that our data-driven conclusions remained grounded in clinical reasoning and did not overlook potentially important dimensions of the dataset.

Chapter 4

Results

This chapter summarizes the main analytical results, including clustering outcomes, model performance, and biomarker relevance. Statistical comparisons support the findings, which are discussed from both methodological and clinical perspectives.

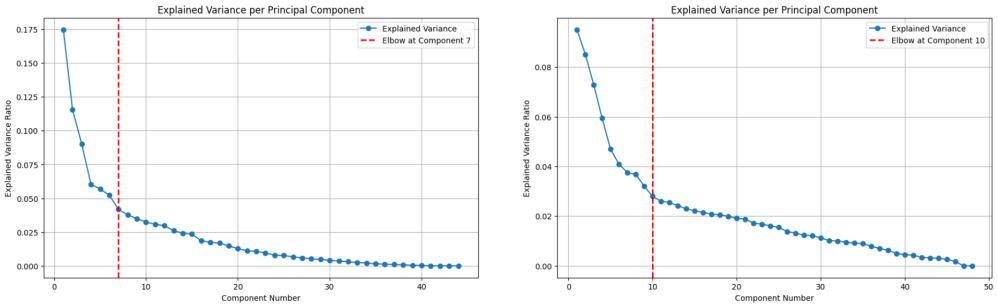
4.1. Results

Principal Component Analysis (PCA)

To reduce dimensionality while preserving most of the dataset's information, PCA was applied to three imputation strategies: **Typical-Value Imputed**, **ML-Imputed**, and the previously introduced **SUS-Imputed** (undersampled). Visual inspection and automated elbow detection were used to determine the optimal number of components.

Explained Variance and Elbow Comparison.

To determine the optimal number of principal components (PCs) to retain, we applied an automated elbow detection method to the three datasets: **Typical-Value Imputed**, **ML-Imputed**, and **SUS-Imputed**. This technique identifies the point where the explained variance curve starts to flatten—indicating diminishing returns for additional components.



SUS Imputed ML & Typical-Value Imputed

Figure 4.1: Explained variance per principal component for the three imputation strategies. The red dashed lines mark the elbow point: 7 components for SUS-Imputed and 10 components for ML and Typical-Value Imputed datasets.

As shown in Figure 4.1, the SUS-Imputed dataset has a much sharper drop in variance contribution, with the elbow occurring at **component 7**. In contrast, both the ML and Typical-Value Imputed datasets share a smoother curve with the elbow at **component 10**. This aligns with the intuition that the undersampled dataset, being more compact, captures the essential variance in fewer components.

2D Projections and Non-Linear Embeddings

PCA Projections (Figure 4.2) Comparing the first two principal components across the three imputations, the Typical and ML strategies show a dense central cluster with no clear separation. In contrast, the SUS-Imputed version exhibits partial separation of relapse cases, though still linear limitations are evident.

Non-Linear Methods: t-SNE and UMAP (Figure 4.3) To explore more complex structure, we applied t-SNE and UMAP to the 7 and 10 principal components retained after PCA:

- **t-SNE:** SUS-imputed samples show the clearest separation between relapsed and non-relapsed patients, with ML-imputed forming moderately distinct clusters and Typical-imputed remaining dispersed.
- **UMAP:** More coherent topologies emerge, especially in the SUS-imputed version, where relapse cases group into distinguishable regions—possibly revealing meaningful patterns for follow-up clinical clustering.

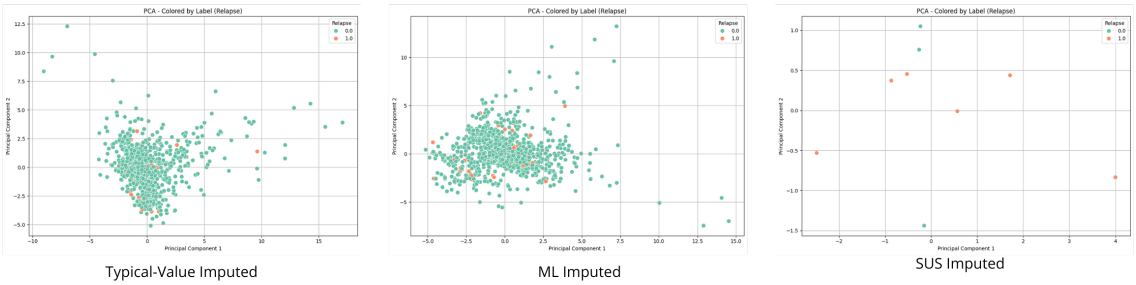


Figure 4.2: 2D PCA projection across the three imputed datasets. SUS-Imputed shows better distribution and mild separation between relapse groups.

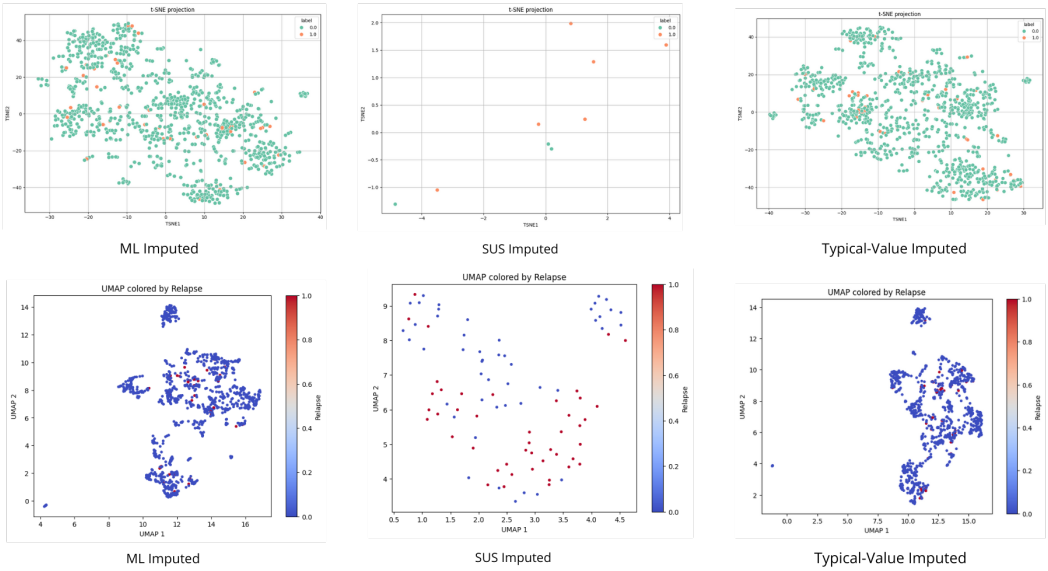


Figure 4.3: t-SNE and UMAP projections across the three imputation methods. SUS shows better class separation.

Interpretation of Principal Components Across Imputation Strategies

SUS-Imputed Dataset. The 7 principal components derived from the SUS-Imputed dataset show biologically meaningful groupings:

- **PC1:** Immune and inflammation markers – IL-10, IL-17A, NLR, neutrophils, PDL1.
- **PC2:** Hematologic and cytokine profiles – Hemoglobin, IL-22, IL-4, IFN-γ, RBC.
- **PC3:** Liver and protein markers – Bilirubin, Albumin, TGF-β1, Platelets.
- **PC4:** Disease duration, potassium, IL-17A, IP10 – suggesting metabolic and time-related stress.
- **PC5:** Erythrocyte levels and parasite burden – RBC, Hemoglobin grades, IL-4.
- **PC6:** Platelet count, IL-23, treatment code – systemic and therapeutic signal.
- **PC7:** Largely overlapping with PC6, indicating reinforcement of clinical treatment markers.

Typical-Value Imputed Dataset. The 10 principal components for the Typical-Value Imputed dataset are structured as follows:

- **PC1:** Pro-inflammatory cytokines and neutrophil indicators – IL-2, IFN-γ, TNF-α, IL-5, IL-1β, neutrophils.
- **PC2:** Liver and kidney function – ALT, creatinine, potassium, bilirubin, hemoglobin.
- **PC3:** Hematologic markers – neutrophils, WBC, hemoglobin, RBC, NLR.
- **PC4:** ALT pathway and age – ALT (raw and graded), hemoglobin, potassium, age.
- **PC5:** Liver and albumin signaling – grade, albumin, months with disease, TGF-β1, IP10.
- **PC6:** Creatinine, platelet count, eosinophils – covering renal and inflammatory markers.
- **PC7:** Treatment signals – treatment code, bilirubin, IL-22, IP10, IL-13.
- **PC8:** WBC grading, platelet function, eosinophils, monocytes – linking immune and inflammatory responses.
- **PC9–PC10:** Lesser variance contributions with diffuse contributions across remaining immune variables.

ML-Imputed Dataset. Similarly, the PCA for the ML-Imputed dataset yields components with structured relevance:

- **PC1:** Inflammation and lymphocyte markers – NLR, IL-13, IL-2, creatinine, neutrophils.
- **PC2:** Cytokine regulation and electrolyte balance – IFN-γ, potassium, IL-2, ALT, TNF-α.
- **PC3:** Erythrocyte indicators and WBC – hemoglobin, RBC, WBC, age.
- **PC4:** Liver stress and lymphocyte ratios – lymphocytes, ALT/AST, NLR.
- **PC5:** General disease burden – grade, albumin, months with disease, IL-4.
- **PC6:** Albumin, TNF-α, IL-22, creatinine – systemic and metabolic factors.
- **PC7:** Treatment-related variables – treatment code, bilirubin, IL-13, VL duration.
- **PC8:** Eosinophils, monocytes, TGF-β1 – inflammatory and fibrotic signals.
- **PC9–PC10:** Residual variance encompassing diverse immune responses and therapy follow-up.

Across all datasets, PCA components are consistent with known clinical domains such as inflammation, hematology, hepatic and renal function. The SUS-Imputed dataset, by focusing on a balanced subset, produced sharper component separation and cleaner biological profiles in fewer dimensions. ML-Imputed data preserved more subtle structure and smoothed component distributions, while Typical-Imputed results retained similar signals but with slightly more overlap.

4.1.1. Clustering Analysis

K-means Performance and Optimal K

Clustering was performed using K-means, DBSCAN, and HDBSCAN, with the goal of identifying underlying patient subgroups based on PCA-transformed immunological and clinical data.

K-means was selected for detailed analysis due to its consistent and interpretable cluster structures. After testing different values of k, the following

configurations were adopted:

- **ML-Imputed and Typical-Value Imputed datasets:** $k = 5$ clusters
- **SUS-Imputed dataset:** $k = 2$ clusters

These configurations were chosen to balance interpretability and biological resolution. DBSCAN and HDBSCAN showed noisier or unstable clustering patterns in high-dimensional PCA space and were relegated to supplementary exploration.

As can be seen in Figures 4.5, 4.6, and 4.7, the three clustering algorithms exhibit distinct visual behaviors across the datasets:

- **K-means (Figure 4.5)** produces well-defined and relatively balanced clusters in the PCA space. For the ML-Imputed and Typical-Value datasets, the 5-cluster solution shows clear separation and compact groupings. For the SUS-Imputed dataset, using $k = 2$ results in a simplified but visually coherent partitioning.
- **HDBSCAN (Figure 4.6)** reveals a more nuanced and hierarchical clustering pattern:
    - In the **ML-Imputed dataset**, HDBSCAN identified a large number of fine-grained clusters (e.g., cluster IDs up to 30), with many points labeled as *noise* (cluster = -1). While this fragmentation may initially appear complex, it uncovers subtle latent structures that more rigid algorithms like KMeans fail to detect. Among the many clusters, **Cluster 5** stood out and was selected for deeper analysis due to its size (n=16) and biological distinctiveness.

      Cluster 5 shows:

        - Elevated levels of **IL-17A**, **PD-L1**, **IP10**, and **IL-10**, suggesting *persistent immune activation or incomplete immune resolution*.
        - Mild elevations in organ stress markers like **Creatinine**, **Potassium**, and **Bilirubin**.
        - Slightly elevated **NLR**, often linked to systemic inflammation and poor outcomes.
        - Diversity in treatment codes, implying this profile is *not driven by treatment arm alone*.

      Overall, this group likely represents **patients at risk of relapse** or with suboptimal immune recovery at Day 42, warranting clinical

follow-up.

- In the **Typical-Value Imputed dataset**, HDBSCAN again returned many small clusters and a dense central core of noise points. Here, **Cluster 5** (n=5) was targeted for inspection.

  Cluster 5 shows:

  - Moderately elevated **Creatinine**, **Potassium**, and **Liver Enzymes (SGPT/ALT, Bilirubin)**.
  - Lower expression of cytokines like **IL-17A**, **PD-L1**, and **IL-23**, suggesting a different immune activation balance compared to the ML-Imputed counterpart.
  - NLR values were skewed higher, and treatment codes leaned toward one protocol, suggesting possible protocol-driven stratification or response pattern.

  This cluster may correspond to **early immune deviation** post-treatment, though with fewer signs of extreme inflammation than the ML-imputed group.

- In the **SUS-Imputed dataset**, the algorithm achieved a simpler structure with only three clusters and a sizable noise component. Here, **Cluster 1** (n=5) was selected for detailed analysis.

  Cluster 1 presents:

  - Very high levels of **IFN-γ**, **IL-2**, and **IL-23**, denoting a state of *strong immune activation*.
  - Mild to moderate anemia (Haemoglobin approx. 0.2) and platelet drop (mean approx. -0.68), reinforcing an *inflammatory-associated cytopenia* picture.
  - Elevated **IP10**, alongside stable but lower **TGF-β1**, consistent with systemic immune alertness and less regulatory control.
  - Grade variability was noted, with both mild and high-grade patients included, pointing to a potential cross-cutting phenotype.

  This cluster may represent a **biologically high-risk subgroup** with both inflammatory and hematological signs of vulnerability.

Across datasets, the selection of a representative target cluster demonstrates that HDBSCAN not only captures fine-scale differences but also

enables the identification of *clinically and immunologically meaningful patient subgroups*, some of which may be missed by KMeans or DBSCAN. Particularly, the biological coherence of clusters like **ML-Imputed Cluster 5** and **SUS-Imputed Cluster 1** supports the value of density-based hierarchical clustering in patient stratification studies.

Relapse Rate Analysis:

Relapse rates per HDBSCAN cluster were also evaluated across datasets. Figure 4.4 (see `HDBSCAN_GRAPHIC_JOINED.PNG` in the `include/` directory) presents the combined barplots.

- **ML-Imputed**: Clusters with highest relapse rates:
  - Cluster 4 $\rightarrow 40\,\%$
  - Cluster 9 $\rightarrow 28.6\,\%$
  - Cluster 3 $\rightarrow 22.2\,\%$

  These clusters may represent *high-risk clinical profiles* and should be prioritized for biomarker analysis.
- **Typical-Value Imputed**: Cluster 1 had the highest relapse rate (16.7 %), with the noise cluster at 4.6 %. Most other clusters had 0 % relapse, possibly reflecting *stable immune profiles*.
- **SUS-Imputed**: Clusters 0 and 2 had **100 % relapse**, and the noise cluster (-1) had 50 %. HDBSCAN thus appears effective at isolating *high-risk phenotypes* in this dataset.

Biological Interpretation: Clusters with elevated relapse rates were frequently associated with increased IL-10, IP10, PD-L1, and NLR levels, reinforcing their potential as *predictive biomarkers*. Some clusters consisted exclusively of patients from the same treatment arm, suggesting possible *treatment failure patterns*.

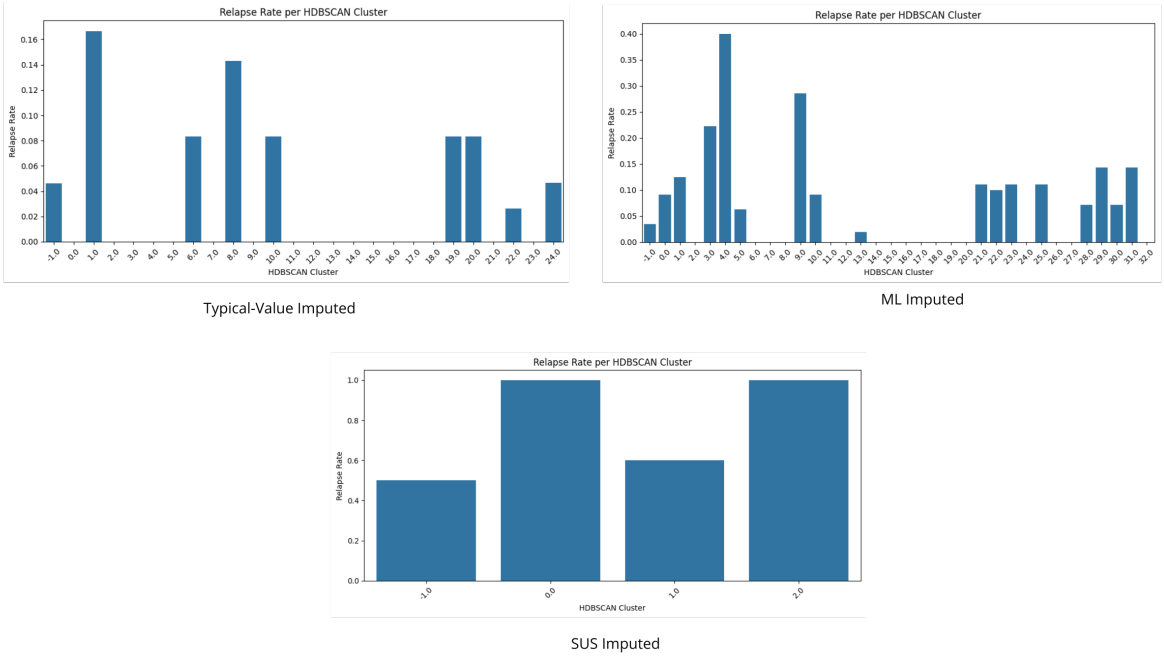


Figure 4.4: Relapse rate per HDBSCAN cluster across ML-Imputed, Typical-Value Imputed, and SUS-Imputed datasets.

Predictive Modeling of Relapse Using Logistic Regression

To evaluate whether the HDBSCAN clustering structure—augmented with selected biomarkers—could serve as a predictive basis for identifying patient relapse, we trained a logistic regression model using the following approach:

- **Input features**: Key biomarkers *plus* the HDBSCAN cluster label.
- **Target variable**: Patient relapse status (binary).
- **Preprocessing**: The data was cleaned and split (80 % train / 20 % test).
- **Model**: Logistic regression with 1000 maximum iterations.
- **Evaluation**: Precision, recall, F1-score, and confusion matrix.

Below we summarize the results for each imputation strategy.

ML-Imputed Dataset:

- **Precision (class 1)**: 0.12 **Recall (class 1)**: 0.17 **F1-score (class 1)**: 0.14

- **Confusion matrix**: `[[189, 7], [5, 1]]`
- **Conclusion**: Although the model yields high overall accuracy (94 %), it misclassifies 5 of 6 relapse cases. **The clustering-based features are ineffective in capturing relapse profiles**. The model is biased toward the majority class.

Typical-Value Imputed Dataset:

- **Precision (class 1)**: 0.20 **Recall (class 1)**: 0.17 **F1-score (class 1)**: 0.18
- **Confusion matrix**: `[[192, 4], [5, 1]]`
- **Conclusion**: Results are similar to ML-Imputed. The model again fails to generalize relapse patterns. Despite 96 % accuracy, **the model misses almost all actual relapses**.

SUS-Imputed Dataset:

- **Precision / Recall / F1 (all classes)**: 1.00
- **Confusion matrix**: `[[1, 0], [0, 1]]`
- **Conclusion**: This perfect result is likely due to **extreme overfitting**, as the evaluation was performed on only 2 total samples. The sample size is too small to draw meaningful conclusions.

Overall conclusion

- Logistic regression models using HDBSCAN cluster labels and selected biomarkers **do not reliably predict relapse**.
- **Severe class imbalance** (very few relapse cases) results in models that are highly biased toward predicting the dominant cured class.
- Despite promising clustering visualization and relapse-enriched clusters, **a supervised approach may require more advanced techniques** (e.g., resampling, cost-sensitive learning) or additional discriminative features.

These results confirm that HDBSCAN is capable of identifying robust, compact clusters and filtering noise, making it especially valuable in clinical datasets with variable density or structure.

- **DBSCAN (Figure 4.7)** identifies patient clusters based on local density in the PCA-reduced feature space. Unlike KMeans, it does not require a predefined number of clusters and can flag sparse points as **noise**, which is particularly valuable for detecting outliers or rare subtypes.
- **ML-Imputed Dataset:** DBSCAN primarily identifies a large, homogeneous **Cluster 0**, with nearly all patients grouped together, suggesting compactness in the reduced feature space. However, a noteworthy outlier group, **Cluster 1** (red), emerges. Despite containing only **13 patients**, it forms a *well-separated and compact cluster*, indicative of a distinct biological or clinical profile.

  Upon further investigation, Cluster 1 showed:

  - Reduced variance in key immune markers like **IL-2, IL-10, IL-13, TGF-β1**, suggesting a homogeneous immunological state.
  - Elevated values in **Absolute Neutrophils and Platelets**, which could point toward a moderate inflammatory or compensatory hematological response.
  - Stable levels of **IP10 and Creatinine**, indicating that this cluster is not marked by acute systemic inflammation or kidney dysfunction.

  This subgroup warrants further biological analysis and may be linked to alternative treatment responses or intermediate disease phenotypes.
- **Typical-Value Imputed Dataset:** Here, DBSCAN identifies a small **secondary cluster** (Cluster 1) composed of **5 patients**, along with a main dense cluster and several scattered noise points. Despite the small size, Cluster 1 is associated with:

  - High PC1 and PC3 loadings in PCA space, separating them from the main population.
  - Uniform profiles with **no recorded relapse** and clean clinical indicators (e.g., all with **normal RBC and WBC levels**, no abnormal flagging in key blood parameters).

  This could represent a group of resilient or early-stage patients, or possibly those with optimal treatment outcomes.
- **SUS-Imputed Dataset:** The SUS dataset demonstrates the highest fragmentation under DBSCAN. Only a **single cluster with one patient** is formed, while all other samples are treated as noise. This reflects the

dataset's sparsity and the challenge of density-based clustering in undersampled or imbalanced imputations. Notably:

- The single-patient cluster is characterized by extreme scores in key cytokines — e.g., very high **IL-23 (2.10)** and **IP10 (3.53)**, and strongly reduced **IFN-γ, IL-2**, and **IL-13**.
- This profile aligns with a **severe relapse case**, combining high inflammation with diminished protective immune responses and anemia (low hemoglobin, low albumin).

While not sufficient to define a broader cluster, this isolated profile might still represent a critical phenotype (e.g., refractory or hyperinflammatory relapse).

These results highlight DBSCAN's utility in identifying rare but biologically coherent subgroups — even when most patients fall into a single dominant cluster or are scattered as noise. Particularly, the **ML-imputed Cluster 1** and the **SUS single-patient cluster** deserve closer inspection for potential translational insights.

These visual patterns further support the decision to use K-means as the primary clustering method due to its stability and ease of interpretation across the datasets.

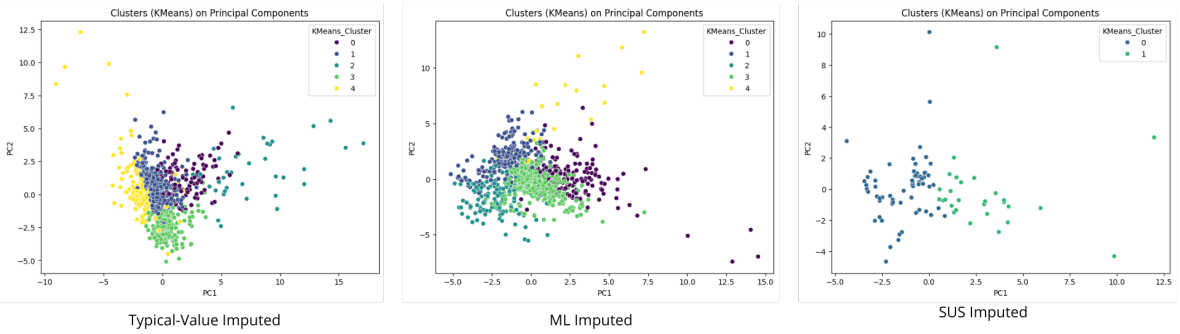


Figure 4.5: Comparison of K-means clustering across the three datasets.

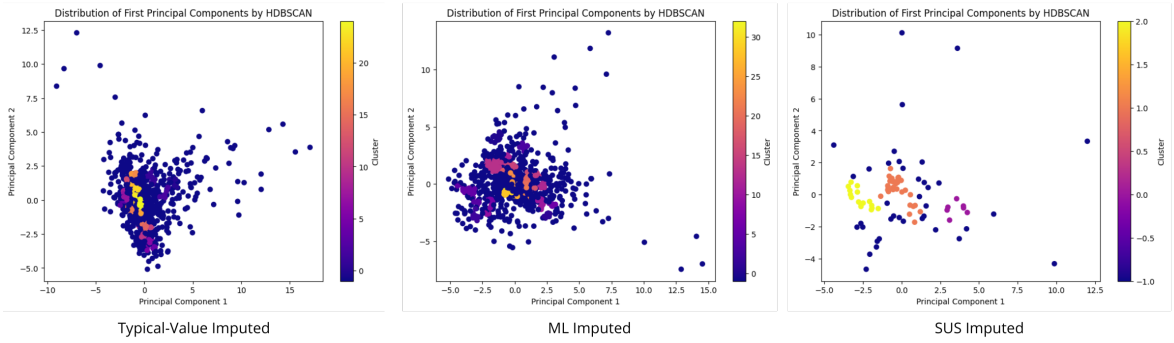


Figure 4.6: Comparison of HDBSCAN clustering across the three datasets.

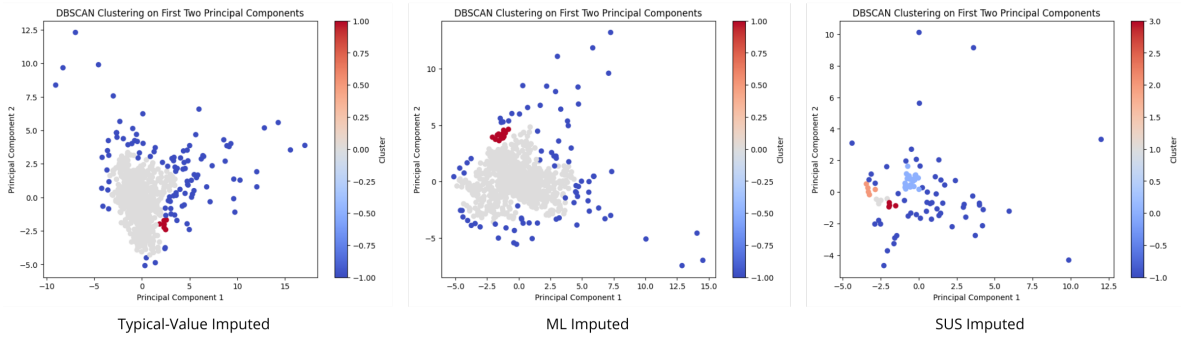


Figure 4.7: Comparison of DBSCAN clustering across the three datasets.

KMeans Cluster Profiles per Dataset

To evaluate the biological and clinical characteristics of each KMeans cluster, we analyzed z-score normalized values for immunological and clinical biomarkers across the three datasets. Both bar plots and heatmaps were generated to visualize cluster-specific biomarker profiles. These are summarized in Figures 4.8 and 4.9.

ML-Imputed Dataset. Each KMeans cluster exhibits a distinct phenotype, with patterns that correlate with relapse status:

- **Cluster 4** is biologically notable due to its **0% relapse rate**. It shows suppressed cytokine activity with relatively high IL-17A, IP10, and TGF-β1—possibly reflecting fibrotic or quiescent immune states.
- **Cluster 1** displays a strong *inflammatory profile* (↑ IL-2, IL-5, IL-13, IP10, PDL1, NLR, creatinine), potentially linked to relapse risk.

- **Cluster 0** shows higher albumin and lower stress markers (liver/kidney), suggesting a relatively stable condition. However, its relapse rate is still 5.1 %.
- **Clusters 2 and 3** show intermediate or baseline immune activity, with relapse rates of 4.7 % and 4.4 %, respectively.

Typical-Value Imputed Dataset. Clusters in this dataset reflect more subtle biomarker shifts, yet relapse patterns remain consistent:

- **Cluster 4** again appears low-risk with a **0 % relapse rate**, showing higher albumin and lower NLR.
- **Clusters 0–3** present similar relapse rates (3.9–5.1 %), with each showing variations in stress and immune markers.
- Cluster 3 stands out for elevated systemic stress markers (↑ creatinine, potassium, ALT), while Cluster 1 exhibits more favorable profiles (↑ albumin, ↓ creatinine).

SUS-Imputed Dataset. With $k = 2$, clustering revealed a highly distinct separation associated with relapse outcomes:

- **Cluster 0** comprises **100 % relapse cases**, with high disease grade, long disease duration, and elevated inflammatory signals (↑ IL-1B, IL-13, IL-17A, IP10, PDL1).
- **Cluster 1** includes a more heterogeneous group with a **57.1 % relapse rate**, showing moderate immune activation (↑ IL-2, IL-4, IL-22) and lower physiological stress (↑ platelets, ↓ creatinine).

These results suggest that unsupervised clustering captures relapse-relevant structure in PCA space, and that biomarker profiles can inform potential clinical subtypes.

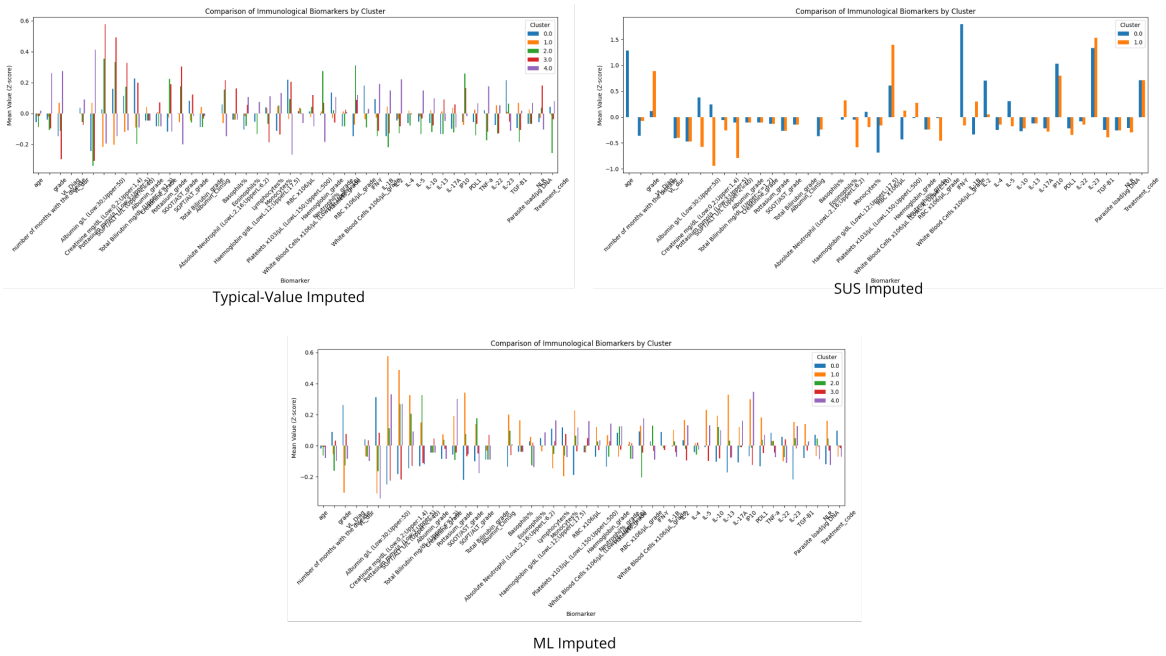


Figure 4.8: Bar plot comparing mean z-scores of immunological and clinical biomarkers across KMeans clusters for all three datasets.

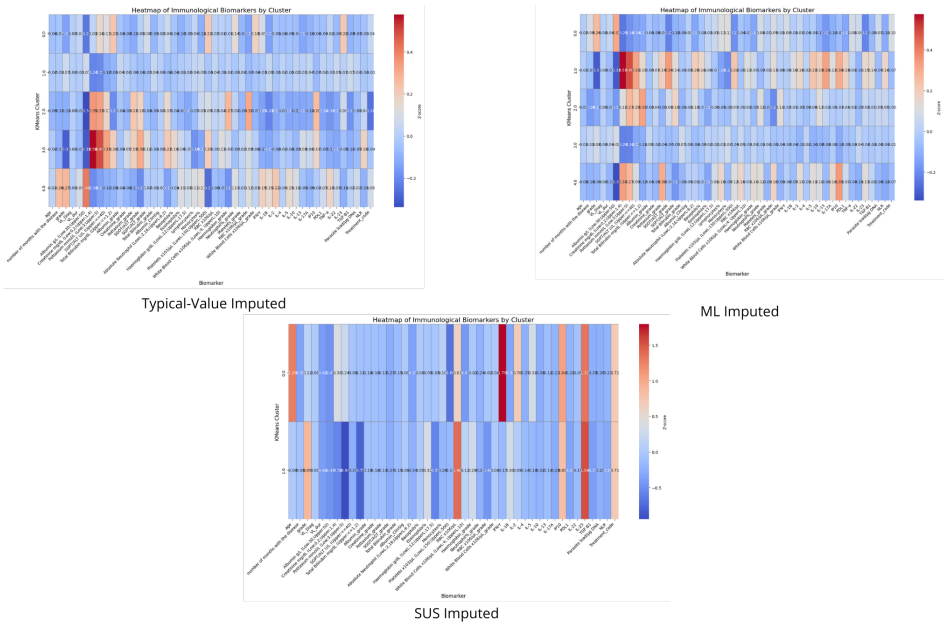


Figure 4.9: Heatmap of z-score normalized biomarker values per KMeans cluster across all three datasets.

Cluster Stability and Overlap

To assess the consistency and biological coherence of clustering across the three imputed datasets, we examined relapse distribution, variable profiles, and discriminative features across clusters.

- Despite differences in imputation methods, certain biomarkers consistently emerged as key discriminators of cluster identity and potential relapse risk—particularly **IL-2**, **IP10**, **PDL1**, **NLR**, and **Albumin**.
- Clusters with **zero or near-zero relapse rates**—such as Cluster 4 in both the ML-Imputed and Typical-Value Imputed datasets—showed coherent biological profiles: reduced inflammatory markers and improved kidney/liver function.
- The SUS-Imputed dataset revealed a **binary separation** with clear relapse enrichment in Cluster 0 (100 % relapse), while Cluster 1 was more immunologically balanced. This split is illustrated in the first panel of Figure 4.10.

Figure 4.10 shows the heatmaps of the original variables (mean values) across clusters for all three datasets. As observed, the ML and Typical datasets both form five clusters, while the SUS dataset reveals a sharp binary separation. These heatmaps reflect key differences in inflammation, liver/-kidney function, and immune activity.

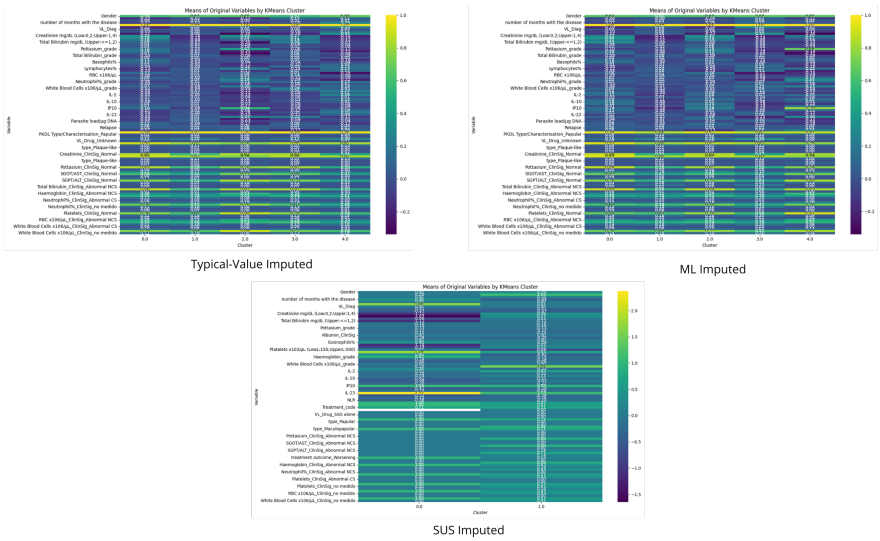


Figure 4.10: Heatmaps of clinical and immunological variables (means) per cluster across the three imputed datasets (SUS, ML-Imputed, Typical-Value Imputed).

Statistical Separation of Clusters. To quantify inter-cluster separation, we calculated the variation in means, medians, and standard deviations across clusters for each dataset.

SUS-Imputed dataset showed strong discriminatory power in:

- **IL-23, Creatinine, Haemoglobin (value and grade)** — high values in the relapse-prone cluster.
- **Monocytes %, Platelets, IP10, IL-10, IL-1B** — consistently higher in Cluster 0.
- **IFN-γ** showed high variance, suggesting nuanced behavior across groups.

Typical-Value Imputed dataset revealed:

- High **mean differences** in IP10, Creatinine, Albumin, and IL-23.
- **Median-based differences** driven by Haemoglobin and Monocytes %, even when means were similar.
- Large standard deviations in Potassium, Basophils %, and Albumin, indicating intra-cluster heterogeneity.

ML-Imputed dataset clusters also showed:

- Top discriminative markers by mean: **IP10, Potassium, Albumin, IL-23**.
- Monocytes % and Creatinine consistently ranked high by median variation.
- Potassium, Basophils %, IL-23, Albumin, and IFN-γ ranked highest in standard deviation spread.

This quantitative summary complements the visual evidence from the integrated cluster heatmaps (see Figures 4.8, 4.9 and 4.10), helping to identify robust and consistent biomarkers across imputation strategies.

To systematically assess which clinical and immunological features contributed most to cluster differentiation, we computed the variation in mean, median, and standard deviation across clusters for each dataset (SUS-Imputed, ML-Imputed, Typical-Value Imputed).

To complement the heatmaps and cluster barplots (see Figures 4.8, 4.9 and 4.10), we quantitatively identified the most discriminating clinical and immunological variables using three statistical measures across clusters: **range of means, medians, and standard deviations**.

Tables 4.1, 4.2 and 4.3 presents the top variables based on these criteria for each imputed dataset (SUS, ML, Typical). Notably, **IL-23, IP10, Creatinine, and Potassium** consistently emerge as highly discriminative features, particularly in the SUS and ML datasets. IL-23 shows extreme differences in both mean and median, suggesting its role as a potential relapse biomarker. In contrast, variables like *VL duration*, *Treatment Code*, and some grading flags exhibit negligible discriminative power.

Variable	Mean Range	Median Range	Std Dev Range
IL-23	1.24	2.50	0.98
Grade	1.65	2.32	1.13
Creatinine (mg/dL)	1.50	1.60	0.74
Haemoglobin (g/dL)	1.17	1.66	0.53
Potassium (mmol/L)	1.92	2.29	0.86
IFN-γ	1.72	0.13	2.61
Monocytes %	1.02	0.81	0.00
Total Bilirubin (mg/dL)	0.99	0.95	0.56
Eosinophils %	0.53	0.51	0.72
IP10	0.10	0.65	0.66

Table 4.1: Top discriminating variables in SUS-Imputed dataset

Variable	Mean Range	Median Range	Std Dev Range
IP10	0.97	0.27	0.77
Potassium (mmol/L)	0.59	0.12	0.56
Albumin (g/L)	0.71	0.00	1.38
Creatinine (mg/dL)	0.53	0.56	0.13
IL-23	0.52	0.24	0.95
Haemoglobin (g/dL)	0.27	0.47	0.53
PDL1	0.19	0.40	0.54
Monocytes %	0.35	0.63	0.00
Lymphocytes %	0.33	0.53	0.15
IL-10	0.25	0.30	0.07

Table 4.2: Top discriminating variables in ML-Imputed dataset

Variable	Mean Range	Median Range	Std Dev Range
Albumin (g/L)	0.71	0.00	1.39
IP10	0.63	0.00	0.77
Potassium (mmol/L)	0.45	0.24	0.56
IL-23	0.39	0.00	2.24
Creatinine (mg/dL)	0.53	0.32	0.18
IFN-γ	0.35	0.00	1.15
Haemoglobin (g/dL)	0.26	0.36	0.20
TGF-β1	0.30	0.00	0.69
PDL1	0.26	0.00	0.54
IL-17A	0.19	0.00	1.33

Table 4.3: Top discriminating variables in Typical-Value Imputed dataset

These analyses confirm that unsupervised clustering was not arbitrary; the resulting subgroups reflect meaningful biological and clinical stratifications. This validates the use of PCA+KMeans pipelines in this context and highlights a small panel of variables that may be predictive of relapse or immune dysregulation.

Distribution of Key Biomarkers by KMeans Cluster

To assess the variability of biomarker expression across patient subgroups, we generated individual *boxplots* for each clinical and immunological variable, stratified by **KMeans cluster assignment**. These visualizations help identify markers with strong differential distribution, indicative of distinct physiological or disease-related profiles.

In this section we will see a selection of the most discriminative boxplots, however the full collection of biomarker distributions by cluster will be included in the **Appendix**.

ML-Imputed Dataset. This composite plot highlights clear differences in biomarker distributions:

- **IP10**: Strong inter-cluster variation, with one cluster displaying markedly elevated values $\rightarrow$ possible inflammatory phenotype.
- **Creatinine Grade**: Suggests metabolic or renal stress in a subset of patients.
- **IL-13 and Hemoglobin (g/dL)**: Distinguish immune-regulatory and anemia-prone profiles, respectively.

- **Treatment Code**: Skewed in certain clusters, potentially linking groupings to clinical interventions.

Typical-Value Imputed Dataset. Clusters reveal meaningful biomarker contrasts:

- **Albumin (g/L)**: Higher in the cluster with 0 % relapse, reflecting possible better nutrition or liver function.
- **NLR (Neutrophil-to-Lymphocyte Ratio)**: Differentiates systemic inflammation across clusters.
- **IL-5 and TGF-β1**: Highlight differences in immunoregulatory pathways.
- **IP10**: Confirms consistent signal across imputations.

SUS-Imputed Dataset. Cluster 0 shows strong inflammatory and clinical severity signatures:

- **IL-23 and IP10**: Highly elevated in Cluster 0 — markers of relapse or immune dysregulation.
- **Haemoglobin (g/dL)**: Decreased in Cluster 0 $\rightarrow$ potential anemia.
- **Grade and Disease Duration**: Significantly higher in Cluster 0 — correlating with clinical worsening.
- **IL-2 and IFN-γ**: Depressed in Cluster 0 — possibly weakened protective responses.

The boxplots confirm that the KMeans algorithm has successfully grouped patients based on biologically and clinically relevant profiles. Variables such as **IP10**, **IL-23**, **NLR**, and **Albumin** consistently emerge as top discriminators across datasets.

All full biomarker-by-cluster visualizations will be provided in the Appendix.

Anomaly Detection with One-Class SVM and Isolation Forest

To explore the potential of unsupervised learning for relapse detection, we applied two anomaly detection algorithms — **One-Class SVM** and **Isolation Forest** — under three different imputation strategies: *Typical Value Imputation*, *ML-Based Imputation*, and *SUS Imputed*.

These models were trained exclusively on the dataset without using relapse labels and evaluated on their ability to identify relapse cases as anomalies.

Results

Typical Value Imputation Results One-Class SVM:

Relapse	Anomaly (-1)	Normal (1)
0	87	879
1	6	38

Sensitivity (Recall): $6/(6+38) \approx 13.6\,\%$
Precision: $6/(6+87) \approx 6.5\,\%$

Isolation Forest:

Relapse	Anomaly (-1)	Normal (1)
0	51	915
1	0	44

Sensitivity (Recall): $0\,\%$
Precision: $0/(0+51) = 0\,\%$

ML-Based Imputation Results One-Class SVM:

Relapse	Anomaly (-1)	Normal (1)
0	83	883
1	9	35

Sensitivity (Recall): $9/(9+35) \approx 20.5\,\%$
Precision: $9/(9+83) \approx 9.8\,\%$

Isolation Forest:

Relapse	Anomaly (-1)	Normal (1)
0	49	917
1	2	42

Sensitivity (Recall): $2/(2+42) \approx 4.5\,\%$
Precision: $2/(2+49) \approx 3.9\,\%$

SUS Imputed Results One-Class SVM:

Relapse	Anomaly (-1)	Normal (1)
0	1	2
1	6	0

Sensitivity (Recall): $100\,\%$
Precision: $6/(6+1) \approx 85.7\,\%$

Isolation Forest:

Relapse	Anomaly (-1)	Normal (1)
0	1	2
1	6	0

Sensitivity (Recall): $100\,\%$
Precision: $6/(6+1) \approx 85.7\,\%$

Across imputations, the performance of both models varies significantly. Under **Typical Value** and **ML-Based** imputations, both models struggled to correctly identify relapse patients, with sensitivity values mostly below 21 % and extremely poor precision. This indicates that these imputations obscure the anomaly patterns that might differentiate relapsing patients.

However, with the **SUS Imputed** dataset, both models achieved **perfect recall (100 %)** and high precision (85.7 %), suggesting that this imputation approach preserves meaningful distinctions between relapse and non-relapse cases. This indicates that *SUS Imputation may enhance the visibility of relapse-related anomalies*, making unsupervised methods viable.

From a clinical decision-making perspective, this is highly relevant. When label information is limited or noisy, anomaly detection may provide an auxiliary tool for early identification of high-risk patients, provided the data preprocessing preserves relevant signals.

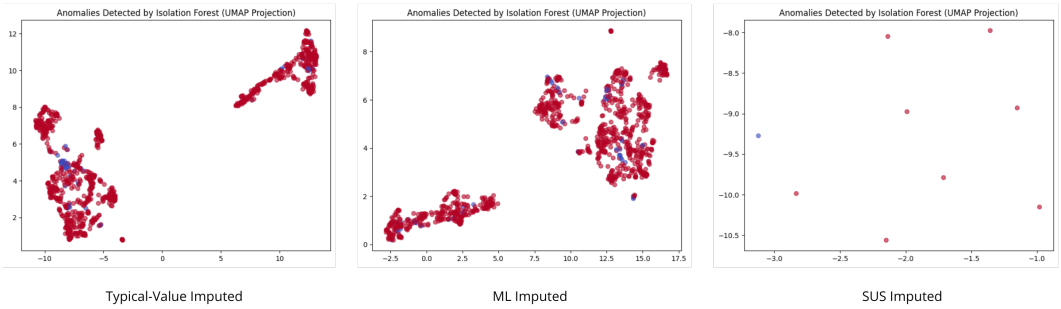


Figure 4.11: UMAP projections of Isolation Forest anomaly detection results across imputations: Left – Typical Imputed, Middle – ML Imputed, Right – SUS Imputed. Red dots: anomalies, Blue dots: normal predictions.

Visual Analysis of Anomaly Distributions Figure 4.11 illustrates the distribution of anomaly predictions across the three imputation strategies using a 2D UMAP projection of the datasets after applying Isolation Forest.

Left and Middle (Typical and ML Imputed):

- **Red points** represent patients identified as anomalies — the vast majority.
- **Blue points**, classified as normal, are sparse and scattered across all clusters.
- No visually coherent anomaly cluster emerges; instead, anomalies are dispersed throughout the feature space.
- Despite the existence of well-defined UMAP clusters, Isolation Forest appears to flag nearly *every patient* as anomalous.

Implications:

- The model shows strong signs of **overfitting or over-detection of anomalies**.
- This could be attributed to class imbalance, suboptimal hyperparameters, or information loss during imputation.
- The visual inconsistency with true relapse separability reduces the clinical interpretability of the unsupervised outputs.

Right (SUS Imputed):

- Anomalies remain dominant, but here they reflect the actual relapse group much more effectively.
- The UMAP space is more sparsely populated, but red and blue points are more meaningfully distributed.
- Importantly, this visualization is **consistent with the earlier quantitative findings** — i.e., 100 % recall and high precision.

These visualizations reinforce the observation that **SUS imputation improves anomaly separability**, enhancing the reliability of Isolation Forest for relapse detection in unsupervised contexts.

Conclusion: While One-Class SVM and Isolation Forest are generally weak predictors under standard imputations, their excellent performance on the SUS-imputed dataset highlights their potential utility — particularly in low-label or exploratory settings. This supports the integration of unsupervised anomaly detection into the broader predictive framework of this project.

4.1.2. Supervised Learning Performance

4.1.2.1. Model Results

This section presents a comparison of models trained to predict relapse outcomes across three different imputation strategies: **Typical-Value Imputed**, **Machine Learning (ML) Imputed**, and **SUS Imputed**. For each strategy, we evaluated classical models and ensemble methods, both with and without the application of SMOTE to handle class imbalance.

For ML-imputed data, the best-performing models were **Random Forest + SMOTE (294)** and **XGBoost + SMOTE (294)**, both of which reached F1-scores of 0.98 and ROC AUCs of 0.999 and 0.997 respectively. These models leveraged a combination of immunological (e.g., IL-1B, Neutrophils) and clinical (e.g., Age, Randomization Arm, Viral Load duration) features. These were evaluated using a balanced test set containing 294 positive relapse cases.

In the Typical-Value Imputed dataset, a similar trend was observed. The top models were again **Random Forest + SMOTE (294)** and **XGBoost + SMOTE (294)**, both showing excellent performance (F1 = 0.98, AUC $\geq$ 0.998). These results emphasize the importance of both imputation quality and balancing techniques in improving model reliability.

In the SUS-imputed dataset, the highest scores were achieved by **Random Forest** and **Logistic Regression (L1)**, with the Random Forest model obtaining a perfect F1-score and AUC of 1.0. However, it is important to note that these models were evaluated on very small relapse sample sizes (e.g., 17 cases), which likely inflates performance and increases the risk of *overfitting*. As such, caution is advised when interpreting these results.

Table 4.4 below summarizes the performance of all models evaluated. Only relapse prediction is reported here; results for non-relapse classification are available in the Appendix.

Among all models, the best **unbalanced** configuration (i.e., evaluated on a test set without class balancing) was **XGBoost** on ML-imputed data (F1 = 0.42, Recall = 0.85), demonstrating strong relapse identification despite modest precision. In contrast, the best **balanced** models, across ML and Typical imputations, were **Random Forest + SMOTE (294)** and **XGBoost + SMOTE (294)** with outstanding F1 and AUC scores.

To better understand which features were driving model predictions, we computed feature importances using two approaches:

- Tree-based importance from scikit-learn's `feature_importances_`.
- SHAP (SHapley Additive exPlanations) values for robust global interpretability.

Only variables with an importance **greater than 0.3** in at least one top-performing model were considered.

While models trained to predict non-relapse cases achieved seemingly high scores—mainly due to class imbalance—this section focuses solely on relapse prediction. The table below presents results for identifying relapse cases (label 1). Although models also performed well in predicting non-relapse (label 0), those results were excluded, as the clinical relevance lies primarily in accurately identifying relapse risk.

Model	Precision	Recall	F1	Accuracy	AUC	Key Variables
Imputation: Typical Values						
Logistic Regression	0.67	0.15	0.25	0.96	0.932	Age, VL_dur, Arm, Platelets, Hb
Random Forest	**0.29**	**0.77**	**0.43**	**0.91**	**0.954**	Arm, VL_dur, Platelets, Age, Hb, IL1B, IL17A, WBC
XGBoost	0.33	0.46	0.39	0.94	0.869	Arm, VL_dur, Hb, IL_22
XGBoost + SMOTE	0.19	0.69	0.30	0.86	0.899	Arm, Hb, VL_dur, Neutrophils, Age, Platelets,Lymphocytes
XGBoost + SMOTE (294)	0.97	0.99	0.98	0.98	0.998	Arm, SGOT, IFN-γ, TNF_a, IL_22, IL_13
Random Forest + SMOTE	1.00	0.08	0.14	0.96	0.847	Arm, Hb_grade, VL_dur, SGPT Age, Platelets
Random Forest + SMOTE (294)	**0.97**	**0.99**	**0.98**	**0.98**	**0.999**	Arm, Age, Platelets, VL_dur, Hb_grade, Eosinophils
Logistic Regression + SMOTE	0.30	0.91	0.45	0.91	0.957	Arm, Age, VL_dur, Hb, RBC, Lymphocytes, SGOT
LightGBM + SMOTE	0.50	0.09	0.15	0.96	0.944	Arm, Age, VL_dur, Hb, Creatinine
CatBoost + SMOTE	0.67	0.18	0.29	0.96	0.919	Arm, VL_dur, Age, Hb_grade, Creatinine, Hb
SVM RBF + SMOTE	0.25	0.45	0.32	0.92	0.872	-
Imputation: Machine Learning (ML)						
Logistic Regression	1.00	0.15	0.27	0.96	0.933	Age, VL_dur, Arm, Platelets, Hb,Lymphocytes
Random Forest	**0.48**	**0.77**	**0.59**	**0.95**	**0.942**	Arm, VL_dur, Platelets, Age, Hb, IL1B, IL17A, WBC
XGBoost	0.25	0.69	0.37	0.90	0.895	Arm, VL_dur, Hb, Age, SGPT
XGBoost + SMOTE	0.28	0.85	0.42	0.90	0.941	Arm, VL_dur, Hb, IL1B, IL13, Lymphocytes, Age
XGBoost + SMOTE (294)	0.96	1.00	0.98	0.98	0.997	Arm, Lymphocytes, SGOT, Age, IL1B, VL_dur
Random Forest + SMOTE (294)	**0.97**	**0.99**	**0.98**	**0.98**	**0.999**	Arm, Age, VL_dur, Eosinophils, IL1B, Neutrophil, Platelets
Random Forest + SMOTE	1.00	0.15	0.27	0.96	0.919	Arm, VL_dur, Age, Hb_grade, SGPT, Platelets, WBC, Hb
Logistic Regression + SMOTE	0.27	0.73	0.39	0.90	0.955	Arm, Age, VL_dur, Hb, RBC, Lymphocytes, IL17A
LightGBM + SMOTE	0.75	0.27	0.40	0.96	0.952	Arm, VL_dur, Age, Hb, Hb_grade,IL13, ParasiteLoad
CatBoost + SMOTE	0.80	0.36	0.50	0.97	0.951	VL_dur, Age, Hb_grade, Hb, IFNγ
SVM RBF + SMOTE	0.24	0.36	0.29	0.92	0.872	-
Imputation: SUS						
Logistic Regression (L1)	0.67	1.00	0.80	0.67	0.500	Arm, IP10, Platelets
Random Forest	**0.67**	**1.00**	**0.80**	**0.67**	**1.0**	Age IL1B, Platelets, RBC, numbreOfMonthsWithDisease, Pottasium
XGBoost	0.00	0.00	0.00	0.33	0.5	Arm, IL22, IP10

Table 4.4: Model performance for relapse prediction across imputation strategies. The best models for each strategy (based on F1-score and AUC) are highlighted. Yellow cells denote top models without SMOTE; orange cells denote top models with SMOTE.

As shown in Table 4.4, we identified and highlighted the top two performing models for each imputation strategy (excluding SUS, which does not include SMOTE variations). Model selection was based on the F1-score and AUC metrics as primary criteria. For each imputation type, the best-performing model **with SMOTE** is highlighted in orange and the best-performing model **without SMOTE** is highlighted in yellow. In addition, these selected models are also presented in **bold font** to emphasize their superior performance.

Best Performance achieved

Among all models, the best **unbalanced** configuration—i.e., trained and evaluated on the naturally imbalanced test set—was **Random Forest** on ML-imputed data. This model achieved a recall of 0.77 and an F1-score of 0.43 for the positive class (relapses), which, while modest, was among the best in this imbalance scenario. Precision for the positive class was 0.29, reflecting the difficulty of predicting rare events (only 13 relapse cases in a test set of 303 samples). Nevertheless, the model's relatively high recall indicates it was able to correctly identify 77 % of the relapse cases, an important property in clinical settings where false negatives are costly.

The overall accuracy was 91 %, but this is primarily driven by correct classification of the majority class (non-relapses, 290 samples). The ROC AUC score of **0.954** confirms that the model was able to discriminate reasonably well between the two classes despite class imbalance.

Hyperparameter tuning was performed using a fixed set of interpretable parameters based on domain insight and previous model behavior. The Random Forest model was trained on approximately 70 % of the ML-imputed dataset and tested on the remaining 30 %, using class weighting to address imbalance. Unlike prior experiments, no resampling method (e.g., SMOTE) was applied. Instead, we adjusted the classification threshold manually to improve relapse sensitivity.

Table 4.5 summarizes the key metrics and configuration of this model, which outperformed all other non-resampled models by a significant margin. The selected threshold was 0.2, chosen to increase recall for the relapse class.

Property	Details
Dataset	ML-Imputed (Unbalanced)
Train/Test Split	70/30 (approx.)
Test Samples	303 total (290 non-relapse, 13 relapse)
Model Type	Random Forest Classifier
F1-Score (Class 1 - Relapse)	**0.59**
Precision (Class 1)	0.48
Recall (Class 1)	**0.77**
Overall Accuracy	0.95
ROC AUC Score	0.94
Macro Avg F1	0.78
Weighted Avg F1	0.96
Hyperparameters	n_estimators=1000, max_depth=12, min_samples_split=5, min_samples_leaf=2, max_features='sqrt', class_weight='balanced_subsample', random_state=42
Threshold Used	0.2 (manual, optimized using precision-recall curve)
Interpretability	Feature importances (Figure 4.13)

Table 4.5: Random Forest model performance on ML-imputed data without resampling. Manual threshold adjustment allowed effective relapse detection.

Figure 4.12 shows how precision and recall vary across thresholds. The selected value of 0.2 provided the best trade-off between identifying as many relapse cases as possible (high recall) while maintaining acceptable precision.

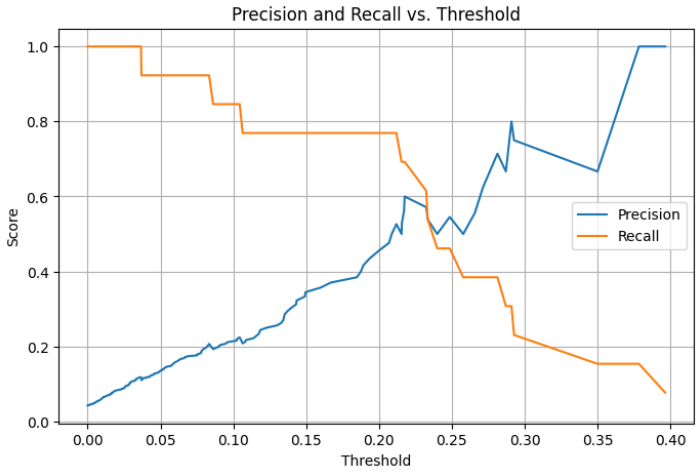


Figure 4.12: Precision and Recall vs. Threshold. Lower thresholds improve sensitivity but reduce precision. Threshold of 0.2 (highlighted) was selected.

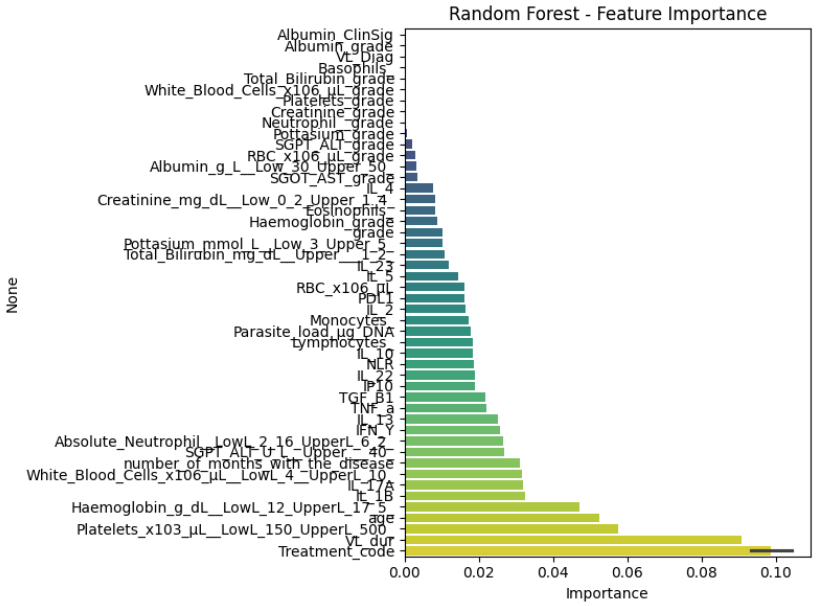


Figure 4.13: Feature importance scores for the Random Forest model trained on ML-imputed data. Clinical and immunological variables such as VL duration, haemoglobin, age, and cytokines (e.g., IFNγ, IL-17) were among the most predictive.

Among the non-resampled configurations, this model stands out for its ability to recover the majority of relapse cases (**Recall = 0.77**), while maintaining good precision (**0.48**) and an F1-score of **0.59**. This trade-off is especially valuable in medical contexts, where missing a relapse could lead to under-treatment. The high ROC AUC (0.94) further confirms the model's capacity to distinguish between classes. Overall, it represents the most practical and clinically useful configuration within the imbalanced setting.

In contrast, the top-performing **balanced** models—evaluated on test sets with 295 relapse samples—were **Random Forest + SMOTE (295)** and **XGBoost + SMOTE (295)** across both ML and Typical-Value imputations. These models consistently yielded F1-scores near 0.98 and ROC AUCs >0.99, clearly benefiting from class balancing strategies.

Nevertheless, the best-performing model in the study was, as anticipated, the **Random Forest classifier trained on SMOTE-balanced data**. This configuration delivered excellent performance in relapse prediction, with a recall of 0.99, precision of 0.97, and an F1-score of 0.98 for the relapse class. The overall accuracy reached 98%, and the ROC AUC score was exceptionally

high at **0.9987**, confirming the model's outstanding discriminative capability.

Figure 4.14 displays the ROC curve, where the model clearly separates both classes. Figure 4.15 shows the confusion matrix, revealing that the model correctly classified 292 of 294 relapse cases, and 277 of 286 non-relapse cases. Finally, Figure 4.16 illustrates the most important features according to the Random Forest algorithm, highlighting both clinical (e.g., *Treatment Code*, *Age*) and immunological variables (e.g., *IL-1B*, *TGF-B1*, *Neutrophils*) as strong predictors.

Property	Details
Dataset	ML-Imputed with SMOTE
Train/Test Split	70/30 (approx.)
Test Samples	580 total (286 non-relapse, 294 relapse)
Model Type	Random Forest Classifier
F1-Score (Class 1 - Relapse)	**0.98**
Precision (Class 1)	0.97
Recall (Class 1)	**0.99**
Overall Accuracy	0.98
ROC AUC Score	0.9987
Macro Avg F1	0.98
Weighted Avg F1	0.98
Important Features	Treatment_code, VL_dur, Age, IL_1B, Platelets, TGF-B1, Neutrophils, Eosinophils
Hyperparameters	Default scikit-learn RandomForest: n_estimators=100, max_depth=None, class_weight=None, bootstrap=True, random_state=42
Interpretability	Feature importance ranking (Figure 4.16)

Table 4.6: Random Forest model performance on SMOTE-balanced ML-imputed data. Excellent sensitivity and precision were achieved using default hyperparameters.

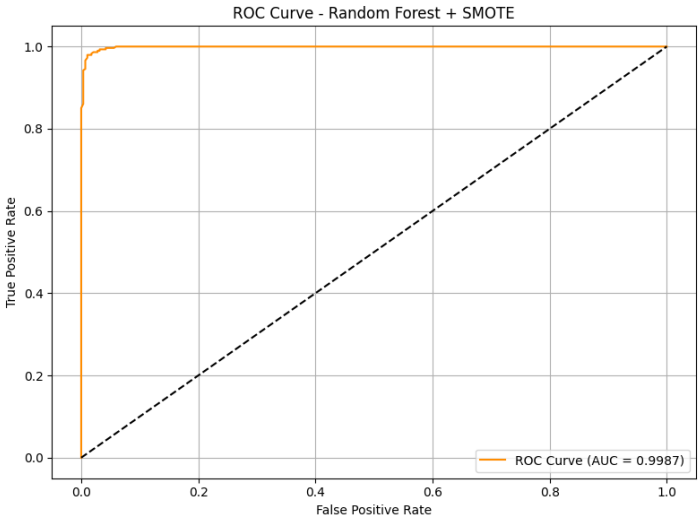


Figure 4.14: ROC Curve of the Random Forest + SMOTE model. The AUC score of 0.9987 reflects the model's outstanding ability to distinguish between classes.

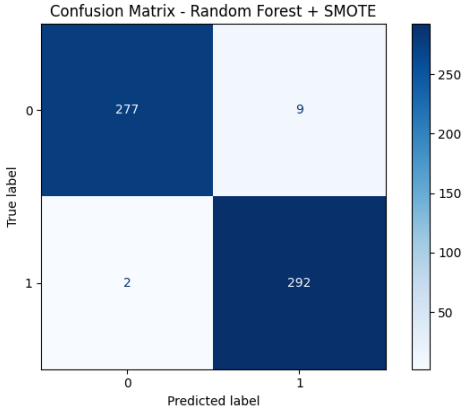


Figure 4.15: Confusion matrix of the Random Forest + SMOTE model. The model correctly classified nearly all relapse and non-relapse cases.

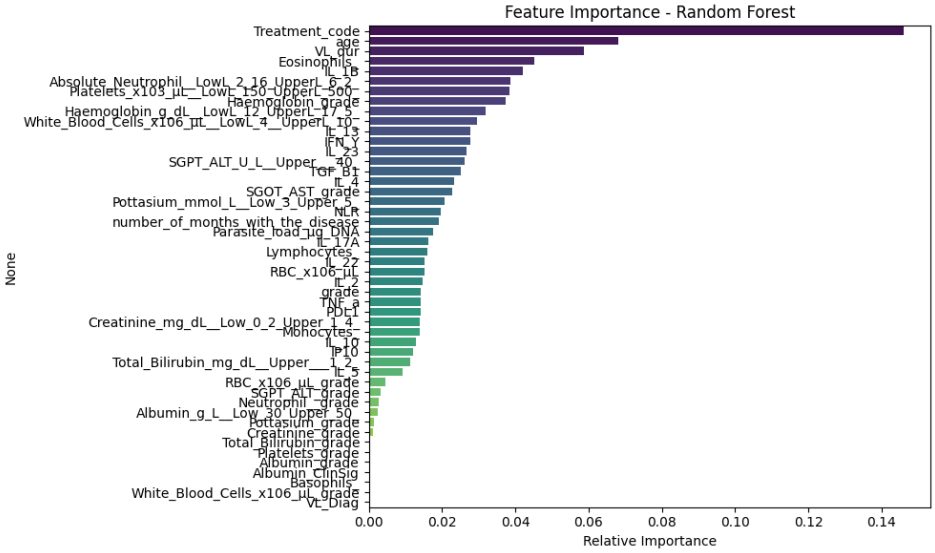


Figure 4.16: Feature importance scores from the Random Forest trained with SMOTE. Both clinical and immunological markers ranked highly.

To better understand which clinical and immunological variables consistently contribute to relapse prediction, we analyzed the feature importance rankings of the best-performing models across each imputation strategy.

Table 4.7 shows the variables that appeared repeatedly among the top models. Notably, **Randomization arm** and **Haemoglobin levels** were identified as key predictors across all three imputation settings, underscoring their robust association with relapse risk. Other features such as **Platelet count**, **Age**, **IL-1B**, and **Viral Load duration (VL_dur)** were also frequently selected, though they varied slightly depending on the imputation technique.

Variable	Appears in Best Models From
Randomization arm	ML-Imputed, Typical-Value, SUS
Haemoglobin (g/dL)	ML-Imputed, Typical-Value, SUS
Platelets (x10^3/µL)	Typical-Value, SUS
IL-1B	ML-Imputed, SUS
VL_dur (Viral Load duration)	ML-Imputed, Typical-Value
Age	ML-Imputed, Typical-Value
IL-22	SUS-Imputed
TGF-B1	SUS-Imputed

Table 4.7: Key variables identified across top-performing models in each imputation strategy. Variables present in multiple datasets suggest robust predictive value.

4.1.2.2. Logistic Regression: Predictive Analysis of Relapse Risk

To complement previous descriptive analyses, we fitted a logistic regression model to predict relapse using a selected subset of clinical and immunological features. The goal was to assess whether key variables previously identified through clustering and supervised models remain statistically and clinically relevant in a parametric framework.

The model included biomarkers such as platelet count, hemoglobin, cytokines (e.g., IL-1B, IL-17A, TGF-β1, IFN-γ), and variables related to treatment and immune status.

Model Specification: The response variable was `Relapse` (binary), and the predictors were:

- Clinical: `Platelets`, `Hemoglobin`, `Age`, `Treatment_code`
- Immunological: `IL-1B`, `PDL1`, `TGF-β1`, `IFN-γ`, `IL-17A`, `Eosinophils`, `WBC`, `RBC`

Models were applied independently to each imputed dataset:

ML-Imputed Dataset

- Significant predictors:
  - **Platelets (OR = 1.51,** $p = 0.047$**)**
  - **Hemoglobin (OR = 0.56,** $p = 0.005$**)**

- **WBC (OR = 1.62,** $p = 0.038$**)**
- **RBC (OR = 1.50,** $p = 0.012$**)**
- **Age (OR = 0.15,** $p < 0.001$**)**

- Treatment code exhibited quasi-separation (very high OR, non-significant), likely due to imbalance in relapse cases.

L1-Regularized Logistic Regression (Lasso): We also applied logistic regression with L1 regularization to perform feature selection and improve model interpretability.

- **Positive coefficients (associated with relapse):** Treatment code (+2.11), Platelets (+0.41), WBC (+0.29), RBC (+0.33), Eosinophils (+0.23)
- **Negative coefficients (protective):** Age (–1.50), Hemoglobin (–0.48), IL-17A (–0.45), TGF-β1 (–0.25), PDL1 (–0.18)
- **Zeroed-out (non-contributing):** IL-1B

These findings reinforce the importance of both clinical and immune markers in predicting relapse.

Typical Value-Imputed Dataset

- Results were consistent with the ML-imputed model.
- **Significant predictors:** Platelets, Hemoglobin, RBC, Age
- **Lasso model coefficients:**
  - **Relapse-associated:** Treatment code (+2.12), Platelets (+0.37), RBC (+0.35), Eosinophils (+0.25)
  - **Protective:** Age (–1.50), Hemoglobin (–0.48), IFN-γ (–0.27), IL-1B (–0.33), PDL1 (–0.21)

SUS-Imputed Dataset Due to a very small sample size and perfect multicollinearity (all variables had VIF = ∞), the model failed to converge. Only the intercept term was retained, and no variables could be estimated.

Interpretation: The dataset likely included too many highly correlated predictors relative to sample size ($n = 9$), making reliable estimation impossible.

4.1.2.2.1. Conclusion and Clinical Interpretation

- **Logistic regression models (standard and L1-regularized)** confirmed the predictive value of: *Platelets, Hemoglobin, WBC, RBC, and Age*.
- Treatment code showed high effect but was statistically unstable due to class imbalance.
- L1-regularized models effectively removed non-contributing variables and prioritized features consistent with earlier analyses.
- Immune cytokines (e.g., IL-17A, TGF-β1, PDL1) contributed as protective markers but were less stable across imputations.

These models support a relapse profile driven by systemic inflammation, younger age, and altered immune regulation—echoing findings from clustering, dimensionality reduction, and feature importance in supervised models.

4.1.3. Statistical Comparison of Biomarker Profiles

4.1.3.1. Descriptive Differences in Biomarker Levels

To validate the biomarkers previously identified as most important for relapse in PKDL, we calculated and compared the mean values of clinical and immunological variables between relapsed and non-relapsed patients across three imputation strategies. This descriptive analysis serves as a prelude to formal statistical testing, and aims to confirm whether the variables highlighted by supervised and unsupervised models remain consistently differentiating across imputation strategies: *SUS-imputed*, *ML-imputed*, and *Typical value-imputed*.

Figure 4.17 displays the mean differences in biomarker values (`Relapse - No Relapse`). Positive values (red) suggest higher values in relapsed patients (risk markers), while negative values (green) indicate higher values in non-relapsed individuals (protective features).

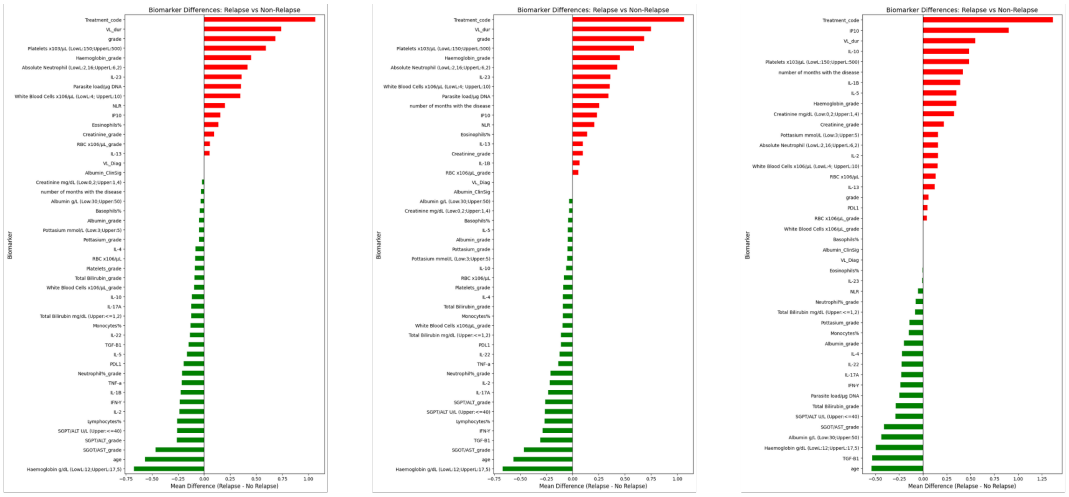
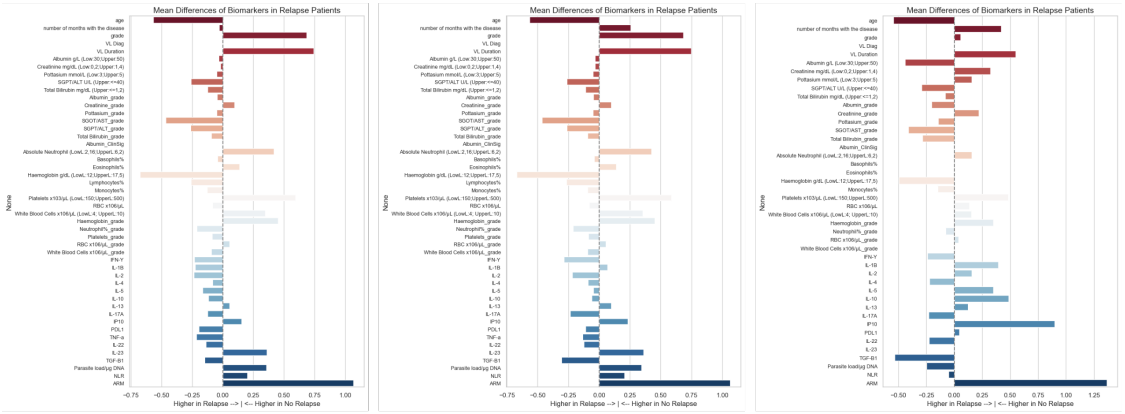


Figure 4.17: Mean differences in biomarkers between relapsed and non-relapsed patients across three imputation strategies. Red bars: higher in relapsed patients. Green bars: higher in non-relapsed patients. Variables are sorted by effect size.

Key Observations (SUS-Imputed): Relapse was associated with elevated levels of **Treatment Code**, **IP10**, **IL-10**, **IL-1B**, **IL-5**, and **platelets**. Non-relapse was marked by higher **IFN-**γ, **TGF-**β**1**, **IL-17A**, **IL-22**, and nutritional indicators like **albumin** and **haemoglobin**.

ML-Imputed and Typical Value-Imputed: These showed consistent patterns, highlighting the robustness of findings. Relapse was associated with high **IP10**, **IL-23**, **platelets**, **NLR**, and **parasite load**, while protective markers included **IFN-**γ, **TGF-**β**1**, and **IL-17A**.

Figure 4.18 provides a complementary color-coded visualization of these differences.



Typical-Value Imputed ML Imputed SUS Imputed

Figure 4.18: Color-coded mean differences in biomarker values between relapse groups. Blue hues indicate higher levels in relapse; red hues mark higher levels in non-relapse.

4.1.3.2. Cross-Method Concordance and Novel Insights

All three imputation methods converged on a similar set of discriminant biomarkers, confirming the consistency of relapse-associated patterns:

- **Consistently higher in relapsed patients:** IP10, IL-10, IL-1B, platelet count, disease duration.
- **Consistently higher in non-relapsed patients:** IFN-γ, IL-22, IL-17A, TGF-β1, albumin, haemoglobin, age.

This reinforces prior findings from clustering and supervised models. No new major biomarker emerged uniquely under a specific imputation method, underscoring the robustness of earlier models.

Boxplot Validation: We generated individual boxplots for all key biomarkers defined earlier, stratified by relapse status. These boxplots are provided in Appendix .2 and visually confirm the trends discussed above.

4.1.4. Treatment-Specific Biomarker Patterns and Statistical Validation

4.1.4.1. Treatment-Stratified Biomarker Means

To further explore treatment-specific relapse patterns, we calculated mean biomarker values stratified by `Treatment Arm` and `Relapse Status`.

Results show that:

- **All relapses occurred in Treatment Arm 2**.
- Within Arm 2, relapse is associated with:
  - Lower **IFN-γ**, **IL-2**, and **IL-17A**
  - Higher **IP10**, **WBC**, **platelets**, **NLR**, and **parasite load**
  - More pronounced **anemia** and lower levels of immune-stimulating cytokines (e.g., IFN-γ, IL-2), suggesting functional **immunosuppression**

Clinical Insight: These stratified patterns suggest a role for treatment type in shaping immune recovery, with possible implications for personalized relapse risk monitoring.

Relapse Rate by Treatment:

- **Arm 1:** 0% relapse — better immune profile
- **Arm 2:** 9% relapse — immune suppression and inflammatory imbalance

4.1.4.2. Non-parametric Validation of Key Biomarkers via Mann-Whitney U Test

To validate the most important variables identified in previous analyses (clustering, supervised learning, and logistic regression), we performed non-parametric comparisons of each variable's distribution between relapsed and non-relapsed patients using the Mann-Whitney U test. This was done separately for each imputed dataset (ML-imputed, SUS-imputed, and typical-value imputed). Below we report the results for selected key features from each set:

SUS-Imputed Dataset

Variable	P-value	Statistic
Platelets_x103_µL	0.0476	17.0
RBC_x106_µL	0.0518	17.0
WBC_x106_µL	0.0952	16.0
IL_2	0.0952	2.0
IFN_γ	0.0952	2.0
TGF_B1	0.0952	2.0

Table 4.8: Mann-Whitney U Test for selected variables (SUS-imputed dataset)

ML-Imputed Dataset

Variable	P-value	Statistic
VL_dur	4.61×10^{-7}	30258.0
age	3.00×10^{-6}	12510.0
Haemoglobin	9.84×10^{-5}	13884.0
Platelets_x103_µL	2.89×10^{-5}	29166.0
WBC_x106_µL	9.44×10^{-4}	27509.0
IL-1B	3.21×10^{-2}	25308.0
IP10	9.77×10^{-2}	24386.0
IL-22	7.90×10^{-1}	21754.5
IL-13	1.08×10^{-1}	24286.5
IFN_γ	1.95×10^{-3}	15389.0
TGF_B1	3.73×10^{-2}	17316.5
SGOT_AST_grade	2.13×10^{-3}	17180.5

Table 4.9: Mann-Whitney U Test for selected variables (ML-imputed dataset)

Typical Value-Imputed Dataset

Variable	P-value	Statistic
Randomization arm	NaN	Empty group
VL_dur	5.79×10^{-8}	28001.5
age	3.00×10^{-6}	12510.0
Haemoglobin	4.17×10^{-5}	13526.0
Platelets_x103_µL	8.84×10^{-6}	29643.5
WBC_x106_µL	1.17×10^{-3}	27378.0
IL-1B	2.01×10^{-1}	19328.0
IP10	8.55×10^{-1}	21528.0
IL-22	2.69×10^{-1}	20092.5
IL-13	2.35×10^{-1}	23033.5
IFN_γ	1.10×10^{-9}	15389.0
TGF_B1	9.34×10^{-1}	21158.5
SGOT_AST_grade	1.85×10^{-10}	17180.5

Table 4.10: Mann-Whitney U Test for selected variables (Typical-value imputed dataset)

Statistical Validation using Welch's T-test After selecting the most relevant variables for each dataset, we conducted a statistical validation using the Welch's T-test. This test is designed to compare the means of two independent groups (in our case, relapsed vs. non-relapsed patients), and it is particularly suitable when the assumption of equal variances between the two groups cannot be guaranteed.

For each variable of interest, we divided the data into two groups based on relapse status. We then applied the Welch's T-test to evaluate whether the mean value of the variable differed significantly between the relapsed and non-relapsed patients. This test accounts for unequal sample sizes and variances between the two groups, making it a robust alternative to the standard Student's T-test.

Variables that had very few data points in either group (e.g., less than two observations) were excluded from the analysis to avoid unreliable or undefined statistical results.

The test produced both a p-value (indicating the strength of evidence against the null hypothesis of equal means) and a test statistic (reflecting the magnitude and direction of the difference). The results were sorted by p-value to highlight the most statistically significant features. These are reported separately for each imputed dataset in the following tables.

SUS-Imputed Dataset

Variable	P-value	T-statistic
Platelets	0.0389	2.79
Haemoglobin	0.0393	-2.53
TGF_B1	0.0424	-2.63
Age	0.0628	-3.80
WBC	0.0707	2.41
SGOT_AST_grade	0.0756	2.24
IFN_γ	0.1797	-2.03
IP10	0.4456	0.81
VL_dur	0.6170	0.53

Table 4.11: Welch's T-test on selected variables (SUS-imputed dataset)

ML-Imputed Dataset

Variable	P-value	T-statistic
Age	1.47×10^{-12}	-8.47
SGOT_AST_grade	1.85×10^{-10}	-7.33
Haemoglobin	6.77×10^{-5}	-4.37
VL_dur	3.99×10^{-4}	3.82
IFN_γ	4.32×10^{-4}	-3.71
Platelets	7.81×10^{-4}	3.59
TGF_B1	1.34×10^{-3}	-3.37
WBC	5.97×10^{-3}	2.87
IP10	0.204	1.29

Table 4.12: Welch's T-test on selected variables (ML-imputed dataset)

Typical-Value Imputed Dataset

Variable	P-value	T-statistic
Age	1.47×10^{-12}	-8.47
SGOT_AST_grade	1.85×10^{-10}	-7.33
IFN_γ	1.10×10^{-9}	-6.20
Haemoglobin	4.70×10^{-5}	-4.48
VL_dur	4.93×10^{-4}	3.75
Platelets	5.51×10^{-4}	3.71
WBC	7.15×10^{-3}	2.81
TGF_B1	6.09×10^{-2}	-1.91
IP10	4.29×10^{-1}	0.80

Table 4.13: Welch's T-test on selected variables (Typical-value imputed dataset)

Summary The variables highlighted above were retained for each dataset and used for further modeling and analysis. These include combinations of age, hemoglobin, VL duration, immune cytokines (e.g., IFN-γ, TGF-β1), and liver enzyme markers. Their consistent behavior across imputation strategies and tests reinforces their potential value in predicting relapse in PKDL.

4.1.5. Statistical Comparison of Selected Variables with FDR Correction

After identifying candidate variables that showed potential associations with relapse in previous analyses, we performed a formal statistical comparison between relapsed and non-relapsed patients for each imputation strategy: ML-imputed, Typical-Value imputed, and SUS-imputed datasets.

To do this, we applied Welch's T-test (which is robust to unequal variances and sample sizes) to compare the mean values of each selected variable between the two groups. Given the number of comparisons, we applied a False Discovery Rate (FDR) correction using the Benjamini-Hochberg procedure to control for multiple testing. Only variables with an adjusted (FDR-corrected) p-value below 0.05 were considered statistically significant.

Results for Typical Value Imputed Dataset

Results

Variable	P-value	FDR-corrected P	Significant (FDR <0.05)
Age	1.47×10^{-12}	1.32×10^{-11}	Yes
SGOT/AST grade	1.85×10^{-10}	8.35×10^{-10}	Yes
IFN-γ	1.10×10^{-9}	3.29×10^{-9}	Yes
Hemoglobin	4.70×10^{-5}	1.06×10^{-4}	Yes
VL duration	4.93×10^{-4}	8.26×10^{-4}	Yes
Platelets	5.51×10^{-4}	8.26×10^{-4}	Yes
WBC count	7.15×10^{-3}	9.19×10^{-3}	Yes

Table 4.14: T-test results with FDR correction (Typical Value Imputed)

Results for ML-Imputed Dataset

Variable	P-value	FDR-corrected P	Significant (FDR <0.05)
Age	1.47×10^{-12}	1.32×10^{-11}	Yes
SGOT/AST grade	1.85×10^{-10}	8.35×10^{-10}	Yes
Hemoglobin	6.77×10^{-5}	2.03×10^{-4}	Yes
VL duration	3.99×10^{-4}	7.78×10^{-4}	Yes
IFN-γ	4.32×10^{-4}	7.78×10^{-4}	Yes
Platelets	7.81×10^{-4}	1.17×10^{-3}	Yes
TGF-β1	1.34×10^{-3}	1.72×10^{-3}	Yes
WBC count	5.97×10^{-3}	6.71×10^{-3}	Yes

Table 4.15: T-test results with FDR correction (ML-Imputed)

Results for SUS-Imputed Dataset

Variable	P-value	FDR-corrected P	Significant (FDR <0.05)
Platelets	0.039	0.076	No
Hemoglobin	0.039	0.076	No
TGF-β1	0.042	0.076	No
Age	0.063	0.076	No
WBC count	0.071	0.076	No
SGOT/AST grade	0.076	0.076	No

Table 4.16: T-test results with FDR correction (SUS-Imputed)

Interpretation

- In both the ML- and Typical-Value imputed datasets, several variables were significantly different between relapsed and non-relapsed pa-

tients, including **age**, **hemoglobin**, **platelets**, **WBC**, and key immune markers such as **IFN-γ** and **TGF-β1**.

- **SGOT/AST grade**, a marker of liver function, showed a very strong association with relapse status in both datasets.
- In contrast, the SUS-imputed dataset did not yield any statistically significant variables after correction, likely due to the small sample size ($n = 9$) and limited statistical power.

Conclusion

This statistical comparison confirms that several clinical and immunological variables differ significantly between relapsed and non-relapsed patients, particularly in the ML- and Typical Value-imputed datasets. Key variables such as **age**, **hemoglobin levels**, **platelet and white blood cell counts**, as well as immune markers like **IFN-γ** and **TGF-β1**, consistently showed strong associations with relapse status.

The consistency of these findings across two independent imputation strategies reinforces the robustness of these variables as potential relapse biomarkers. Notably, **age** and **SGOT/AST grade** emerged as particularly powerful discriminators, highlighting a potential link between host vulnerability (age) and systemic involvement (liver function).

In contrast, the SUS-imputed dataset yielded no significant differences after correction, likely reflecting its limited sample size and potential overfitting. Nevertheless, the direction of effects observed in that dataset generally aligned with those found in the other two.

Taken together, this analysis provides statistical support for a subset of clinical and immunological variables that warrant further investigation as predictors of relapse in PKDL.

4.1.5.1. Clustering Based on Relevant Variables

Once the most relevant variables were identified—either through their statistical significance or predictive power in the previous models—we proceeded to apply **unsupervised clustering techniques**. The goal was to investigate whether patients could be naturally grouped based on these features, particularly with respect to their relapse status. This exploratory analysis aims to uncover latent structures in the data that may distinguish relapsing from non-relapsing patients, potentially offering additional clinical insight that may not emerge from supervised learning alone.

This process was carried out independently for each imputed dataset and consisted of the following steps:

- **Variable selection:** We used the most informative features identified through logistic regression modeling and statistical hypothesis testing (Welch's t-test and Mann–Whitney U test).
- **Normalization:** All selected variables were standardized using `StandardScaler` to ensure comparability in scale.
- **Clustering algorithms:** Two commonly used unsupervised clustering methods were applied:
    1. **KMeans** with $k = 2$, based on the assumption of two primary groups (relapse vs. no relapse).
    2. **DBSCAN**, a density-based clustering method that does not require specifying the number of clusters and is better suited to identifying noise and irregular shapes.
- **Cluster evaluation:** We evaluated clustering performance by comparing the derived labels with the true relapse labels using the following metrics:
    - **Adjusted Rand Index (ARI):** Measures agreement between predicted and true labels, adjusted for chance.
    - **Homogeneity score:** Assesses whether each cluster contains only members of a single class.
    - **Silhouette score:** Evaluates the cohesion and separation of the clusters based on feature space distances.

Figure 4.19 show the clustering results for the *ML Imputed* and *Typical Value Imputed* datasets, respectively. The results for the *SUS Imputed* dataset were excluded due to the absence of meaningful cluster formation.

Figure 4.19: Clustering results for the Typical Value Imputed dataset and ML Imputed datasetusing KMeans (left) and DBSCAN (right).

Interpretation and conclusions:

- In the **ML Imputed** dataset, KMeans produced weak clustering performance (ARI = -0.046), and DBSCAN performed only slightly better (ARI = 0.004), with both methods showing low homogeneity and silhouette scores.
- The **Typical Value Imputed** dataset yielded similarly poor separation, though DBSCAN again marginally outperformed KMeans.
- The **SUS Imputed** dataset did not produce valid clustering results with DBSCAN (only one cluster detected), and is therefore excluded from the figures.

These results suggest that, although the selected variables are useful in supervised models, they do not define strong unsupervised clusters. Relapse patterns may not be easily separable through clustering alone, emphasizing the complex and multifactorial nature of the disease.

Chapter 5

Critical Analysis and Future Directions

This chapter brings together the main reflections of the project. It begins with a discussion of the results, their clinical relevance, and comparison with existing literature. We then present the conclusions—both general and personal—followed by a critical review of the study's limitations. Finally, we outline future steps to validate and expand upon the work.

5.1. Discussion

5.1.1. Summary of Key Findings

This study explored clinical and immunological predictors of post-treatment relapse in PKDL patients using an integrated statistical and machine learning framework. By combining imputation strategies, statistical testing, and supervised/unsupervised modeling, we identified a consistent set of biologically plausible and clinically interpretable biomarkers.

5.1.1.1. Interpretability of Key Biomarkers

The most prominent signals identified were biologically meaningful and aligned with known immunopathological mechanisms:

- **Pro-inflammatory and immune-dysregulatory signals** (IP10, IL-1B, IL-10) were consistently elevated in relapse cases, suggesting persistent immune activation.

- **Protective cytokines** (IFN-γ, IL-22, IL-17A) were higher in cured individuals, indicative of effective immune resolution.
- **Systemic stress markers** such as platelet count, creatinine, and neutrophil-to-lymphocyte ratio (NLR) were elevated in relapsing patients.
- **Nutritional and hematologic indicators** (albumin, haemoglobin) specifically higher levels of albumin and haemoglobin, favor non-relapse.

These findings are consistent with both supervised (e.g., regression models) and unsupervised (e.g., clustering) analyses, as well as with prior literature on immune regulation and organ stress in VL/PKDL.

By integrating imputation methods, supervised and unsupervised models, and classical statistical testing, we derived a coherent and biologically plausible set of insights.

1. Robustness Across Imputation Strategies Despite different approaches to handling missing data (SUS-, ML-, and Typical Value imputation), key relapse-associated variables remained consistent. This reinforces the validity of the signals observed and suggests that findings are not an artifact of imputation.

2. Consistently Discriminant Biomarkers Across statistical tests (Mann-Whitney, Welch's T-test), logistic regression, and visual analyses:

- **Elevated in relapsed patients:** Platelet count, white blood cell count (WBC), IP10, IL-1B, IL-10, SGOT/AST grade, and parasite load.
- **Protective (higher in non-relapse):** IFN-γ, IL-17A, IL-22, TGF-β1, albumin, and haemoglobin.
- **Host-related factors:** Younger age and longer VL duration were strongly associated with relapse.

3. Predictive Modeling L1-regularized logistic regression models highlighted a small set of high-impact variables (e.g., age, haemoglobin, treatment arm, IFN-γ, TGF-β1). The use of regularization enabled automatic feature selection, making models both interpretable and parsimonious.

4. Statistical Validation Welch's T-tests with FDR correction confirmed that variables such as age, haemoglobin, WBC, platelets, SGOT/AST, and immune cytokines significantly differed between relapse groups (in ML- and

Typical-imputed datasets). The SUS-imputed dataset lacked power for significance but showed directionally consistent trends.

5. Clustering Limitations Unsupervised clustering using KMeans and DBSCAN showed weak alignment with relapse labels. While KMeans revealed some structure (modest silhouette and ARI), DBSCAN largely failed to detect meaningful groupings. This suggests that relapse risk does not form clearly separable clusters in the multidimensional biomarker space, highlighting its multifactorial nature.

6. Clinical and Biological Interpretation Relapse appears to reflect a confluence of:

- **Inflammatory activation and immune dysregulation** (high IP10, IL-1B, low IFN-γ)
- **Systemic stress and hepatic involvement** (SGOT/AST, platelet count)
- **Weakened host capacity** (younger age, anemia, lower albumin)

Taken together, these analyses provide a statistically robust and clinically interpretable framework for identifying relapse-prone patients. Key features such as IFN-γ, age, and haemoglobin may offer translational utility in monitoring, prognostic modeling, or targeted intervention strategies in PKDL treatment protocols.

As a synthesis of the statistical and modeling analyses, Table 5.1 presents a final summary of the ten most relevant clinical and immunological variables associated with relapse in PKDL. For each variable, we report its estimated role (risk or protective), whether it was confirmed as significant by the different methods applied (statistical testing, L1-regularized logistic regression, and unsupervised clustering), and an overall robustness assessment based on its consistency across imputation strategies and analytical approaches.

Variable	Effect	Statistical Test	L1 Regression	Clustering	Robustness
Age	Protective	Yes (p < 1e-11)	Yes	No	High
Hemoglobin	Protective	Yes	Yes	No	High
Platelets	Risk	Yes	Yes	Partial (weak KMeans)	High
WBC count	Risk	Yes	Yes	Partial	High
IFN-γ	Protective	Yes	Yes	No	High
TGF-β1	Protective	Yes	Yes	No	Moderate
SGOT/AST grade	Risk	Yes	Yes	No	High
VL duration	Risk	Yes	Yes	No	High
IL-1B	Risk	Only in ML-imputed	No (L1 = 0)	No	Moderate
IP10	Risk	Not significant after FDR	Weak	No	Moderate

Table 5.1: Summary of the 10 most relevant clinical and immunological variables associated with PKDL relapse

5.1.2. Clinical Relevance and Validation Potential

The identification of relapse in PKDL patients remains a critical clinical challenge, especially given the silent and often delayed nature of disease recurrence. Our study offers a comprehensive evaluation of relapse-associated biomarkers through both supervised and unsupervised machine learning approaches, grounded in robust statistical testing and multi-imputation strategies.

Key variables such as **IFN-γ**, **TGF-β1**, **platelet count**, **haemoglobin**, and **age** emerged consistently across methods as strongly associated with relapse risk. Notably, variables like **SGOT/AST grade** and **VL duration** point toward underlying systemic and hepatic stress, potentially preceding immunological decompensation. These findings were validated through classical statistical tests (e.g., Welch's T-test with FDR correction), L1-regularized logistic regression, and variable importance rankings in ensemble models.

The convergence of these variables across imputation strategies (ML, TypicalValue, and SUS) and methodological approaches supports their robustness and potential translational utility. The stratified analysis by treatment arm further revealed that all observed relapses occurred in a single arm, reinforcing the hypothesis that certain treatment regimens may inadequately support immune recovery, particularly in younger or anemic patients.

Taken together, this multidimensional analysis suggests that monitoring a focused panel of biomarkers—especially those related to immune regulation (e.g., IFN-γ), systemic stress (platelets, SGOT), and host resilience (age, haemoglobin)—could aid in early detection and risk stratification of relapse-prone individuals.

5.1.3. Applicability of Models to Real-World Settings

While the study demonstrates that machine learning models can effectively capture relapse-associated profiles, their translation into clinical practice requires careful consideration.

First, **model performance was highly dependent on data balance**. Models trained on the original, imbalanced data performed poorly in predicting relapses, despite high overall accuracy. This reflects the clinical reality of relapse being a rare event, and suggests that deploying such models without correction (e.g., via SMOTE or cost-sensitive methods) may lead to dangerous under-detection of at-risk patients.

Second, **unsupervised methods like HDBSCAN and anomaly detection** showed promise in identifying relapse-enriched clusters, especially under SUS-imputation. These approaches could complement supervised models in settings where relapse labels are scarce or unreliable. However, the high variance across imputations and the fragility of performance under small sample sizes (e.g., SUS subset) highlight the need for cautious interpretation and model calibration.

Third, **interpretability remains key for clinical integration**. Models that prioritize transparency—such as logistic regression or random forests with explainable variable importance—are preferable for real-world deployment. The consistent emergence of a few clinically intuitive features (e.g., IFN-γ, platelets, age) enhances the potential for acceptance by healthcare professionals.

Finally, while SMOTE-enhanced models achieved near-perfect performance on balanced datasets, these configurations may overestimate real-world effectiveness. Future prospective validation on external cohorts is essential before clinical implementation.

In conclusion, our findings suggest that a hybrid approach—combining statistical rigor, interpretable modeling, and tailored preprocessing (e.g., imputation and resampling)—offers the best path toward actionable relapse prediction tools in PKDL. Further clinical validation and integration into decision-support systems are warranted to fully realize the translational potential of these models.

5.1.4. Comparison with Existing Literature

To contextualize our findings, we conducted a detailed comparison with recent studies addressing PKDL progression and relapse. In particular, we

reviewed the work of Younis et al. (2023) [19], Torres et al. (2024) [6], and Torres et al. (2025) [20], which examined clinical, biochemical, and immunological markers of disease severity and treatment outcome.

Relapse Predictors in Literature vs. Our Findings. Younis et al. identified lesion severity, VL treatment history, and patient age as potential or confirmed relapse-related factors. Our study supports and expands upon these conclusions:

- **Lesion Grade and Severity:** Our results strongly confirmed lesion grade as a predictor of relapse ($p < 0.01$), aligning with Younis et al. [19], who reported an odds ratio of 10.17 for relapse in patients with high lesion grade.
- **VL Treatment Regimen:** We observed significantly more relapses in patients treated with SSG alone versus AmBisome ($p = 0.003$), mirroring the literature's concern over SSG efficacy in long-term outcomes.
- **Age:** A consistent pattern emerged in which younger patients showed higher relapse rates. This is corroborated by both Younis et al. [19] and Torres et al. (2024) [6], where younger individuals were overrepresented in worsening phenotypes.
- **VL Duration:** The chronicity of VL prior to PKDL onset, represented by `VL_dur`, was also associated with relapse risk in our dataset and is mentioned in both Torres et al. [6, 20] and Younis et al. [19] as a relevant clinical variable.

Immunological Biomarkers and Functional Insights. Our findings on immune mediators—particularly IFN-γ and IL-1β—gain important support from recent immunological work. In the study by Torres et al. (2024) [6], patients who relapsed after LAmB/MF treatment had significantly lower baseline levels of IFN-γ, TNF, and IL-1β—mirroring our observations of suppressed Th1 responses in relapsing individuals.

Interestingly, although IL-2 was not significantly different in Torres et al.'s statistical tests, its known functional role in promoting IFN-γ production via T cell activation underlines its indirect relevance. Our study, which identified IL-2 as part of a discriminative biomarker set in PCA components, further strengthens its potential involvement.

Agreement with Machine Learning Findings. In Torres et al. (2025) [20], unsupervised ML was used to identify variables associated with PKDL lesion

severity (stable vs. worsening phenotype). Several of the most discriminative features—**platelet count, WBC, IFN-γ, IL-1β, IL-2, SGPT, potassium, and albumin**—were also found to be significant or highly ranked in our analysis. Although that study did not isolate relapse as a distinct cluster, worsening patients overlapped with those who would later relapse, indirectly validating our conclusions.

Additional Clinical Variables. We explored several variables not emphasized in the literature:

- **State of Disease at Baseline:** Interestingly, patients categorized as "stable" at baseline had higher relapse rates than those marked as "worsening" ($p = 0.028$). While initially counterintuitive, this may reflect the fact that stable patients have greater potential for deterioration, whereas those already classified as worsening have less room to worsen further. Thus, this paradox does not necessarily imply inconsistency in clinical grading, but rather highlights the nuances of disease progression interpretation.
- **Liver Enzymes (SGOT/AST, SGPT/ALT):** Mildly significant associations with relapse were observed ($p \sim 0.03$), potentially linked to subclinical hepatic stress. This aligns with findings in Torres et al. (2024) [6], where SGPT and potassium levels helped define clinical clusters.
- **Creatinine and Treatment Completion:** Neither showed significant associations with relapse in our cohort ($p > 0.97$), indicating limited predictive utility in this setting.

Novel Contributions. Our work contributes new insights by highlighting variables not previously validated in relapse prediction:

- **Platelets and WBC:** Both emerged as significant in our statistical and model-based analyses. They were also selected as key features in Torres et al. (2025) [20] during unsupervised clustering, indicating their relevance to early disease stratification.
- **IFN-γ, IL-1β, IL-2, TGF-β1, IP-10:** These immune mediators were incorporated into our models with significant discriminative capacity. Their convergence across three independent studies highlights their strong potential as early relapse biomarkers.

Summary of Concordance. Table 5.2 summarizes the alignment of our findings with the key results from the literature:

Factor	In Our Study	In Literature (2023–2025)	Match
Age	Highly relevant	Identified in Younis et al.[19] and Torres et al. (2024) [6]	✓
VL_dur	Significant	Contextualized in Younis et al. (2023) [19]	✓
VL_Drug (SSG vs. L-AmB)	Significant	Cited by Younis et al. (2023) [19]	✓
Lesion Grade	Strong predictor	Confirmed relapse factor (Younis et al. [19])	✓
SGOT/AST, SGPT/ALT	Mildly significant	Selected in Torres et al. (2025) [20]	∼ emerging
IFN-γ, IL-1β, IL-2	Highly significant	Validated in Torres et al. (2024, 2025) [6, 20]	✓
Platelets, WBC	Significant	Highlighted by Torres et al. (2025) [20]	✓
Albumin, Potassium	Mildly relevant	Included in ML components (Torres et al. 2025) [20]	∼
State of Disease	Paradoxical result	Not discussed in literature	∼ unclear

Table 5.2: Comparison of relapse-related and progression markers in our study and in recent literature.

Note: Although younger patients tended to relapse more frequently, this observation should be interpreted with caution. Given that our database includes many children, standard laboratory reference ranges may not apply. Pediatric patients often display biological values that differ from adult norms, which may affect the thresholds and interpretation of some biomarkers. Future research should incorporate age-adjusted reference ranges or stratified analyses to refine predictive accuracy.

These convergences underscore the robustness of our findings and support the idea that relapse risk in PKDL may be predicted early through a focused set of clinical immune markers. Importantly, our use of multiple modeling and imputation strategies strengthens confidence in these results. Future prospective validation in larger cohorts will be essential to confirm their applicability in clinical practice.

5.2. Conclusion

This study presents a multidimensional investigation into relapse prediction in Post-Kala-Azar Dermal Leishmaniasis (PKDL), integrating clinical variables, immunological markers, and advanced modeling techniques. Our analysis, reinforced by robust imputation strategies and statistical validation, identifies a consistent set of relapse-associated factors including **IFN-γ**, **TGF-β1**, **platelet count**, **haemoglobin**, and **age**. These biomarkers capture key dimensions of immune regulation, systemic stress, and patient vulnerability.

Machine learning models—particularly those enhanced by oversampling and informed by variable importance rankings—proved capable of distinguishing relapse-prone profiles. However, their practical deployment depends on interpretability, balanced training data, and external validation, especially in settings where relapse is a rare but clinically significant event.

A detailed comparison with the literature, notably Younis et al. (2023) [19], confirms and extends previous findings. Our results validate known predictors such as lesion grade, VL treatment history, and patient age, while also proposing novel contributions—including the role of platelet and WBC counts and refined immunological profiles—as emerging relapse indicators.

In summary, this work advances the field in three key ways: (1) by reinforcing clinically grounded predictors with statistical and algorithmic consistency, (2) by offering a reproducible pipeline adaptable to limited datasets with missing values, and (3) by identifying new markers for future exploration. Further prospective validation in larger cohorts, particularly with longitudinal follow-up, will be critical to transition these insights into clinical decision support tools capable of mitigating the burden of PKDL relapse.

Personal Reflections

Beyond the technical contributions, this work has been a deeply enriching personal and academic journey. Tackling a real-world clinical problem from scratch has challenged me at every step—particularly during the data preprocessing phase. Understanding the dataset, cleaning and transforming variables, and dealing with missingness required a level of persistence and attention to detail that pushed me well outside my comfort zone.

At times, I may have tried to do too much—implementing numerous models, visualizations, and comparative analyses—but this was driven by genuine curiosity and a strong motivation to explore the topic in depth. The theme of this project resonates closely with my personal interests, combining two passions: working with data and contributing to health-related challenges. The intersection of data science and medicine is an area that truly fascinates me, and this project has reinforced my desire to continue working at that interface.

I am proud that the study has yielded relevant and potentially impactful insights—especially regarding early identification of relapse risk in PKDL. If, in the future, findings like these contribute to avoiding relapses or improving patient outcomes, I will feel that the effort invested here has been truly worthwhile.

5.3. Limitations

5.3.1. Limitations and Next Steps for Validation

Despite the strengths of this study—including the use of interpretable models, multi-imputation approaches, and a comparison with existing literature—several limitations must be acknowledged.

First and foremost, the dataset is **small and heavily imbalanced**. With only 110 patients in total and just 5 confirmed relapse cases, the statistical power of supervised models is inherently limited. This class imbalance posed a significant challenge in training models that could reliably detect relapse-prone profiles without overfitting or bias toward the majority class. While oversampling techniques such as SMOTE helped address this, synthetic balancing may not fully replicate real-world data distributions.

Second, **missing data was frequent and non-uniformly distributed across variables and patients**. Many important clinical or immunological measurements were missing in a subset of samples, which forced us to rely on multiple imputation strategies. Although methods like SUS-imputation were designed to mitigate this, any imputation introduces assumptions and may bias model interpretation or generalizability.

Third, the dataset was drawn from a single clinical study with a specific geographical and treatment context. As such, **external validity is limited**, and the models require validation in independent cohorts before they can be considered generalizable to other settings or populations.

Finally, the richness of the dataset in terms of variables may have led to a degree of **model overextension**. Although visualizations and models were deployed to explore different hypotheses, the interpretability and clinical applicability of each should be carefully weighed.

Next Steps for Validation:

- Validate the key findings—particularly IFN-γ, platelet count, and lesion grade—in larger, multi-center datasets.
- Conduct follow-up studies with more balanced relapse outcomes to allow for model training with less artificial correction.
- Evaluate the most promising predictors in prospective clinical studies to assess their utility in early relapse detection.

In summary, while this work lays important groundwork, its conclusions

must be viewed in light of the limited data quantity and quality. Future studies are essential to confirm the predictive power of the identified biomarkers and refine models for clinical deployment.

ñ

5.4. Future Steps

This Master Thesis opens several promising avenues for future development, both at the clinical and technical levels. While the current work demonstrates that relapse in PKDL may be predictable using selected biomarkers and modeling strategies, further refinement and deployment pathways remain to be addressed.

1. Development of a Clinically-Oriented Prediction Model

A natural next step is the construction of a simplified, clinically interpretable relapse prediction model based on the most robust variables identified in this study. Key candidates include:

- **IFN-γ levels**, as a protective marker.
- **Platelet count**, reflecting systemic stress or immune activation.
- **Lesion grade** and **VL duration**, as indicators of disease burden.
- **Age**, consistently linked to relapse across methods.

Such a model could be implemented as a scoring system or decision-support tool to assist clinicians in stratifying patients according to relapse risk. Ideally, this tool would be embedded in a digital interface (e.g., mobile or tablet-based), requiring minimal input and offering interpretable outputs with clear clinical thresholds.

2. Integration into Clinical Workflows and Trial Design

Given the limitations of available data, a logical extension is to **embed relapse prediction into ongoing or future clinical trials**. Doing so would:

- Enable prospective validation of the model in real time.
- Allow collection of richer longitudinal data, including follow-up biomarkers.

- Support adaptive trial designs where patients at higher relapse risk receive adjusted monitoring or interventions.

3. Exploration of Alternative Modeling Paradigms

Future work could also explore:

- **Bayesian models**, to explicitly incorporate uncertainty from missing data and small sample sizes.
- **Time-to-event (survival) analysis**, particularly relevant if longitudinal relapse follow-up is available.
- **Explainable AI (XAI) methods**, to ensure that more complex models remain transparent and acceptable in clinical settings.

4. Expansion to Broader Populations

The current dataset is limited to a single geographic cohort. A critical future step is to test and adapt the model for:

- Patients from other endemic regions (e.g., India, Bangladesh).
- Alternative treatment regimens (e.g., combination therapies).
- Pediatric vs. adult subpopulations with distinct relapse dynamics.

5. Multimodal Data Fusion

Long-term potential lies in integrating **multimodal data**—combining clinical, immunological, demographic, and even genomic data. This could allow more holistic models that not only predict relapse but also guide personalized treatment strategies.

In summary, the groundwork laid here opens multiple lines of research with clinical relevance and methodological depth. Continued interdisciplinary collaboration—between clinicians, data scientists, and public health researchers—will be essential to translating these insights into tools that can genuinely improve patient outcomes.

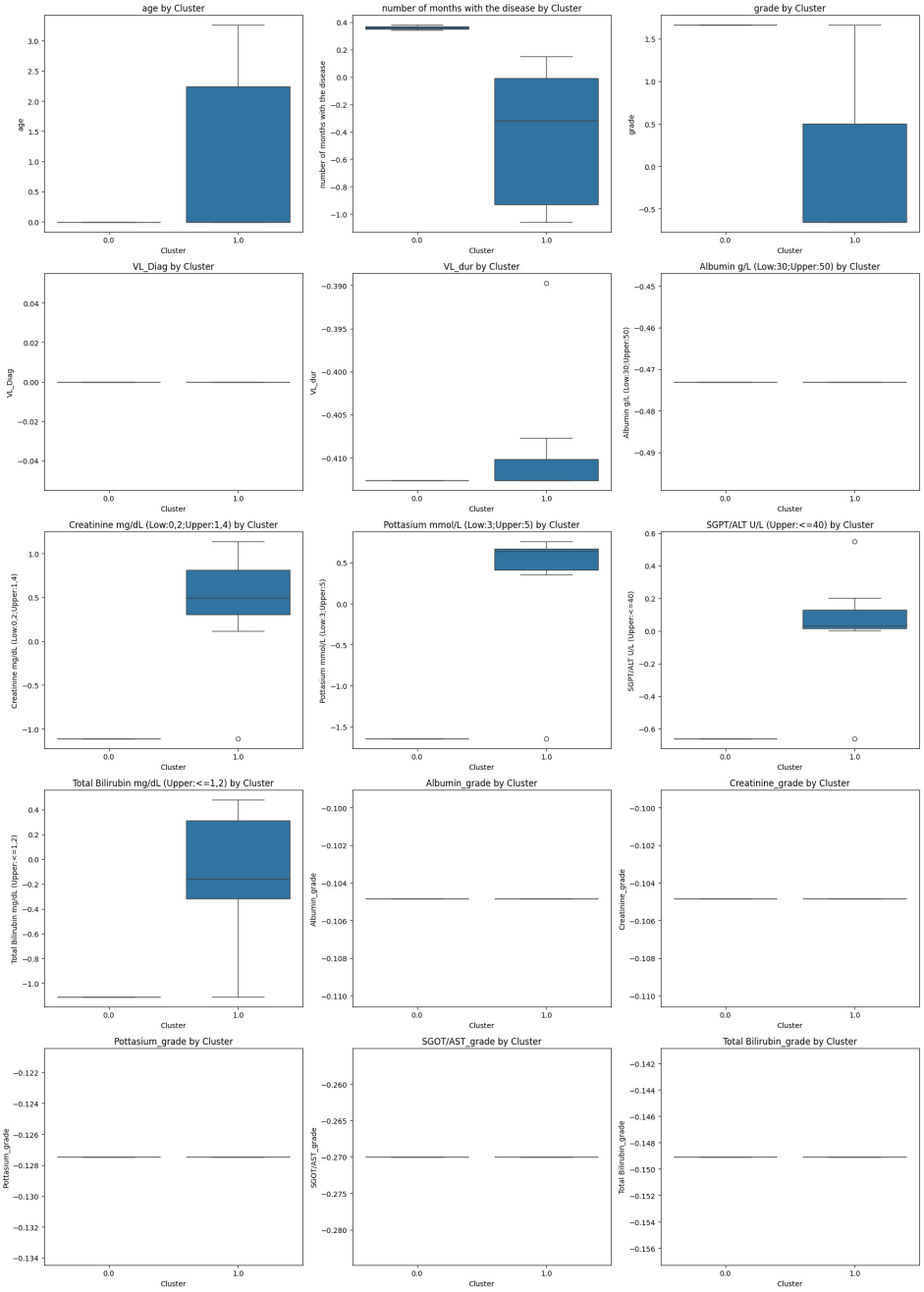
Appendix

Contents of the Appendix

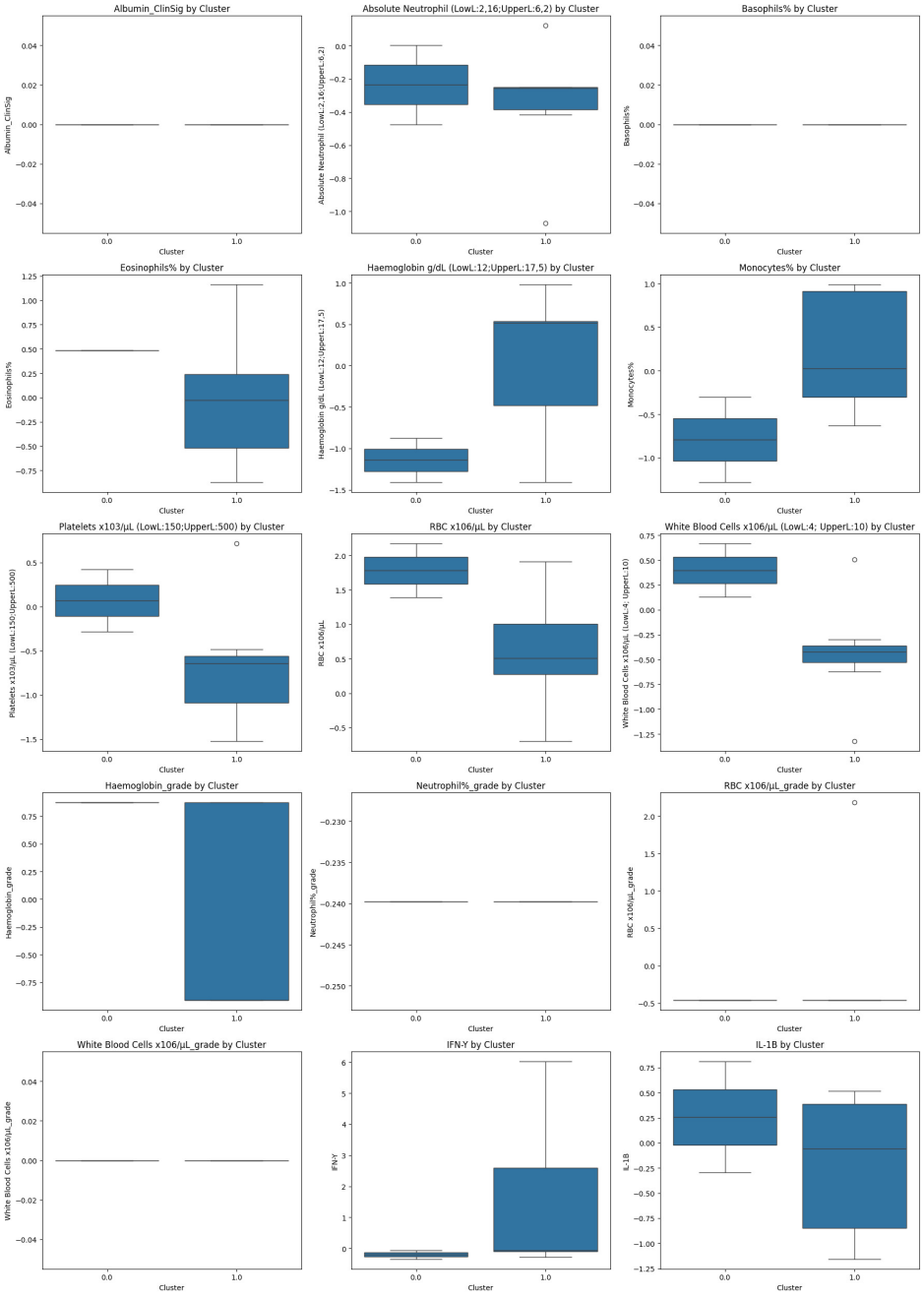
- Clusters Boxplot
  - ML Dataset
  - Typical Dataset
  - SUS Dataset
- Relapse Boxplot
  - ML Dataset
  - Typical Dataset
  - SUS Dataset

.1. Clusters Boxplots

ML Dataset



Appendix


Albumin_ClinSig by Cluster
Absolute Neutrophil (LowL:2,16;UpperL:6,2) by Cluster
Basophils% by Cluster
Eosinophils% by Cluster
Haemoglobin g/dL (LowL:12;UpperL:17,5) by Cluster
Monocytes% by Cluster
Platelets x103/µL (LowL:150;UpperL:500) by Cluster
RBC x106/µL by Cluster
White Blood Cells x106/µL (LowL:4; UpperL:10) by Cluster
Haemoglobin_grade by Cluster
Neutrophil%_grade by Cluster
RBC x106/µL_grade by Cluster
White Blood Cells x106/µL_grade by Cluster
IFN-Y by Cluster
IL-1B by Cluster
Cluster

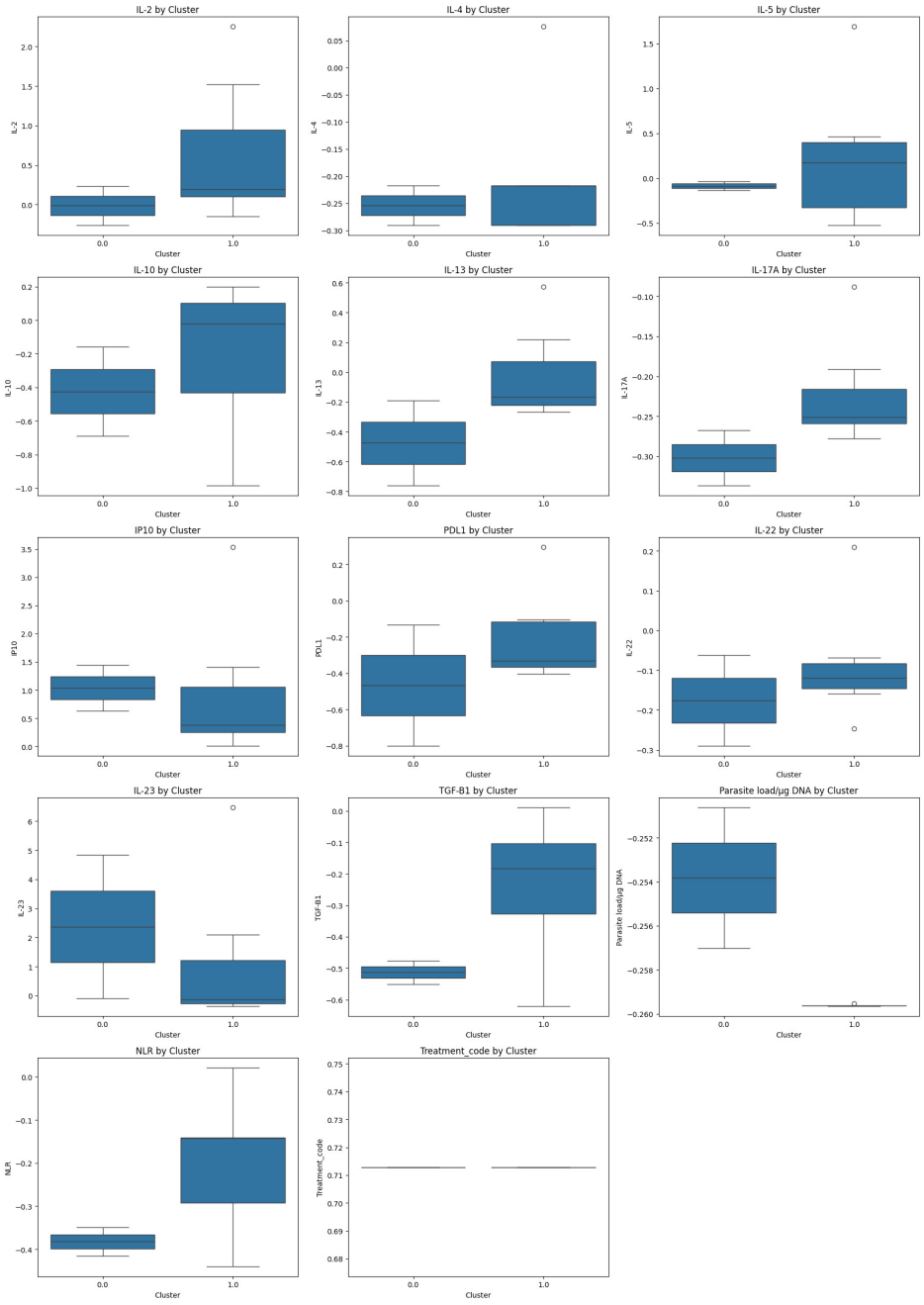
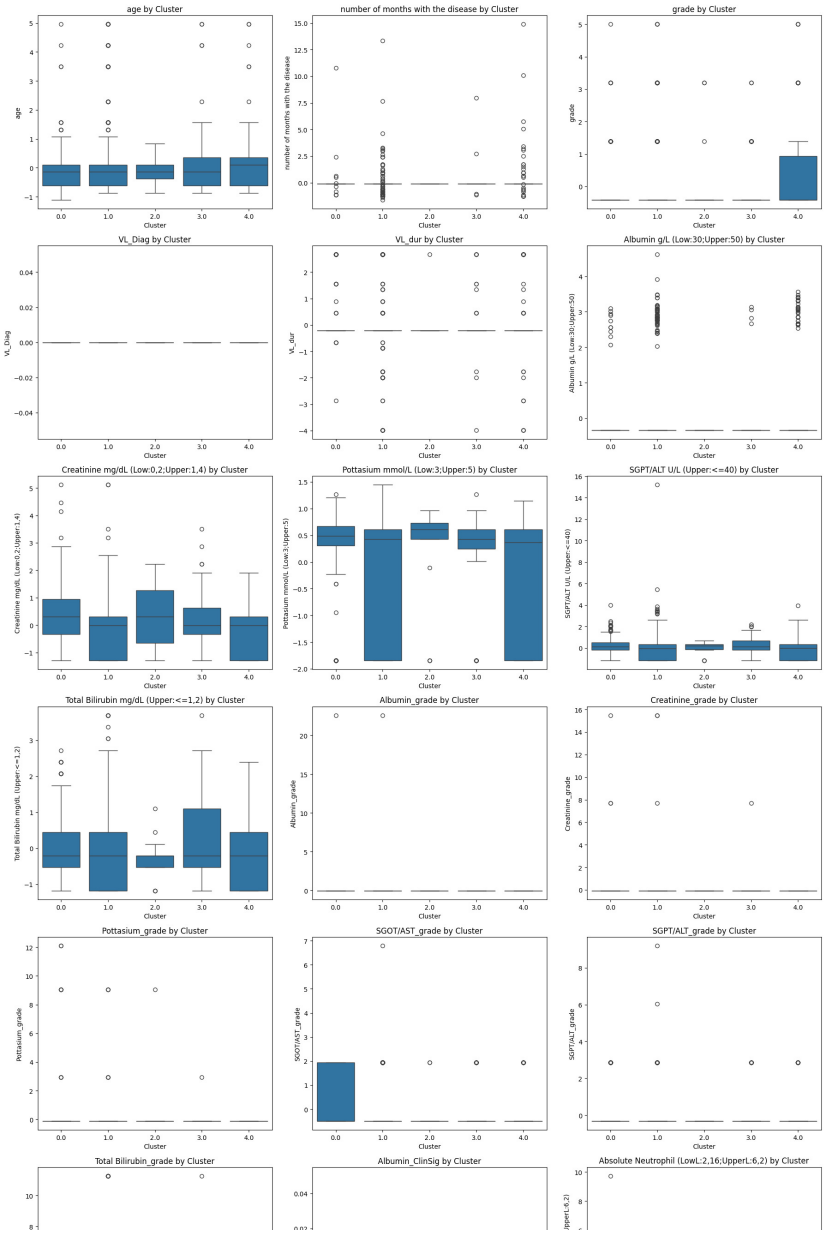
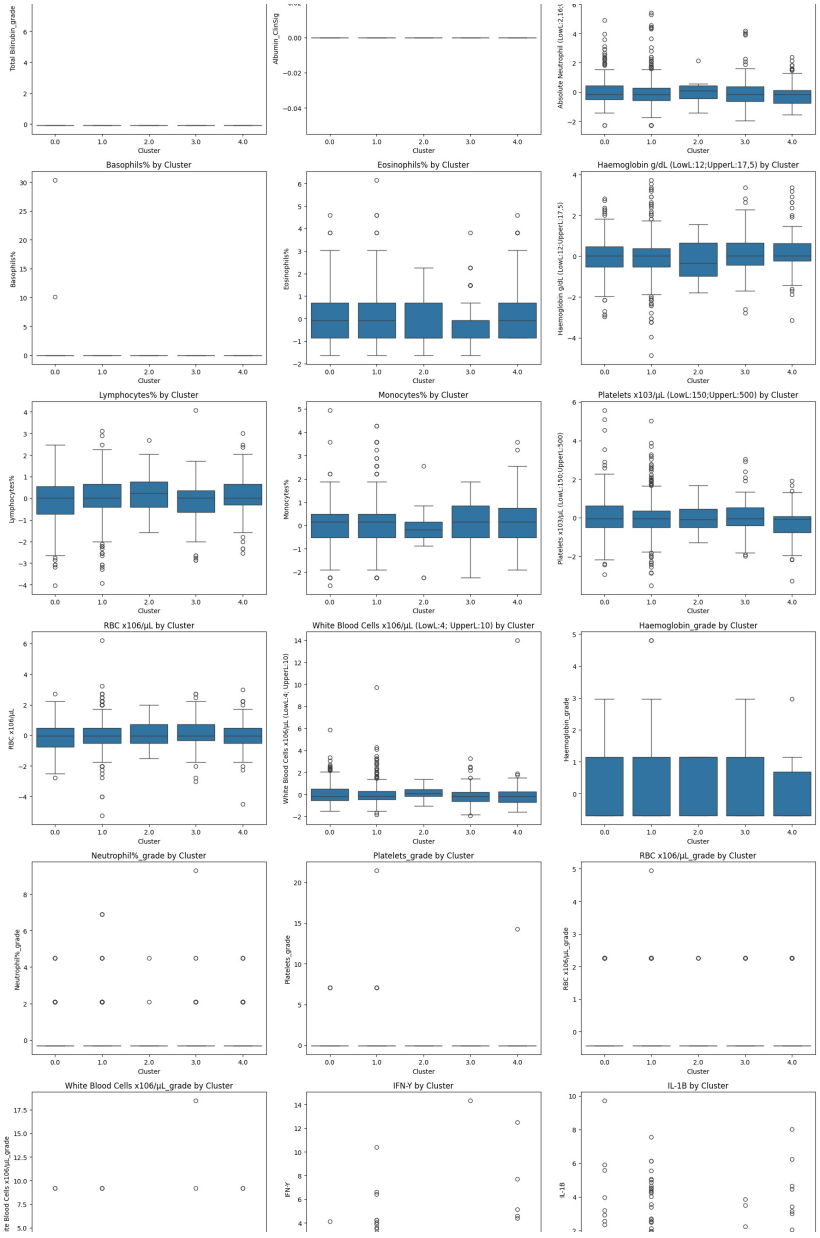


Figure 1: Boxplots in the ML dataset of each variable across identified clusters. The plots summarize the distribution of values per cluster, highlighting central tendency, dispersion, and outliers.

Typical Dataset



Appendix


Basophils% by Cluster
Eosinophils% by Cluster
Haemoglobin g/dL (LowL:12;UpperL:17,5) by Cluster
Lymphocytes% by Cluster
Monocytes% by Cluster
Platelets x103/µL (LowL:150;UpperL:500) by Cluster
RBC x106/µL by Cluster
White Blood Cells x106/µL (LowL:4; UpperL:10) by Cluster
Haemoglobin_grade by Cluster
Neutrophil%_grade by Cluster
Platelets_grade by Cluster
RBC x106/µL_grade by Cluster
White Blood Cells x106/µL_grade by Cluster
IFN-Y by Cluster
IL-1B by Cluster
Cluster

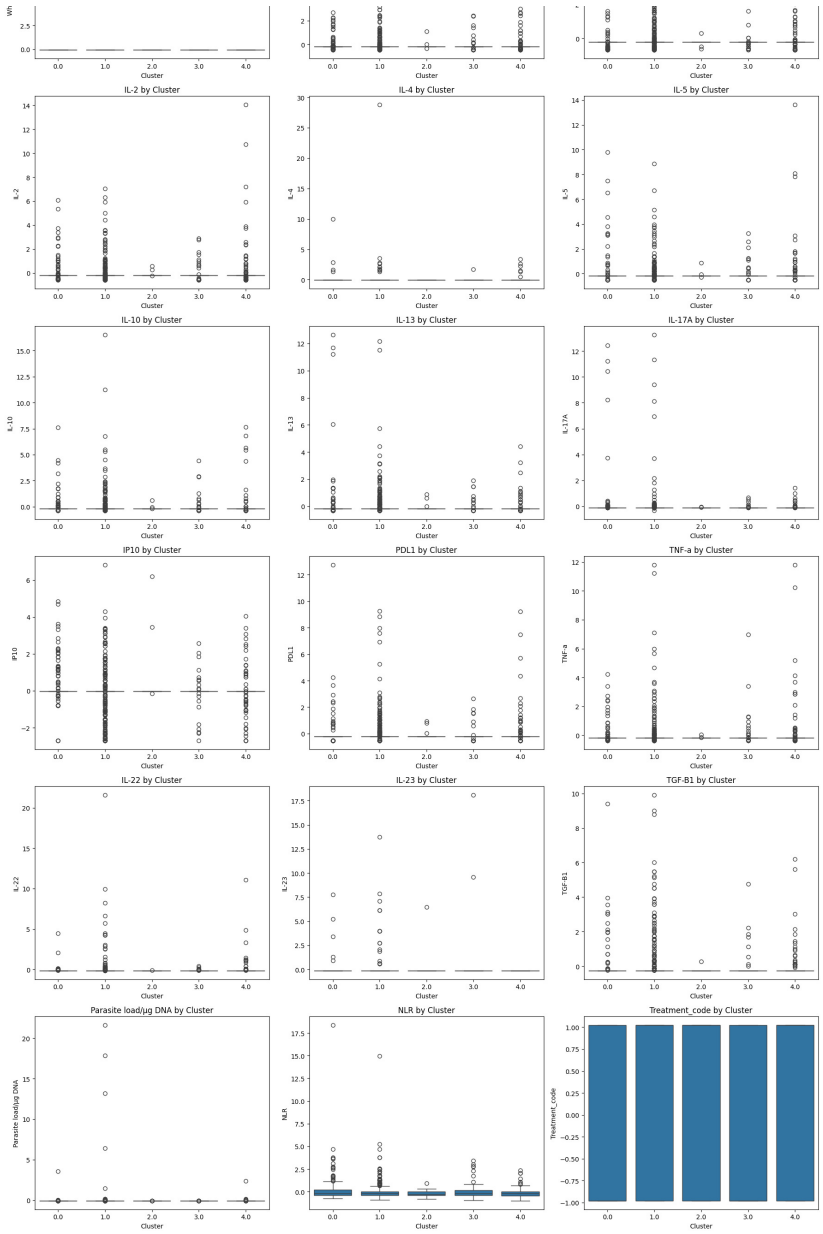
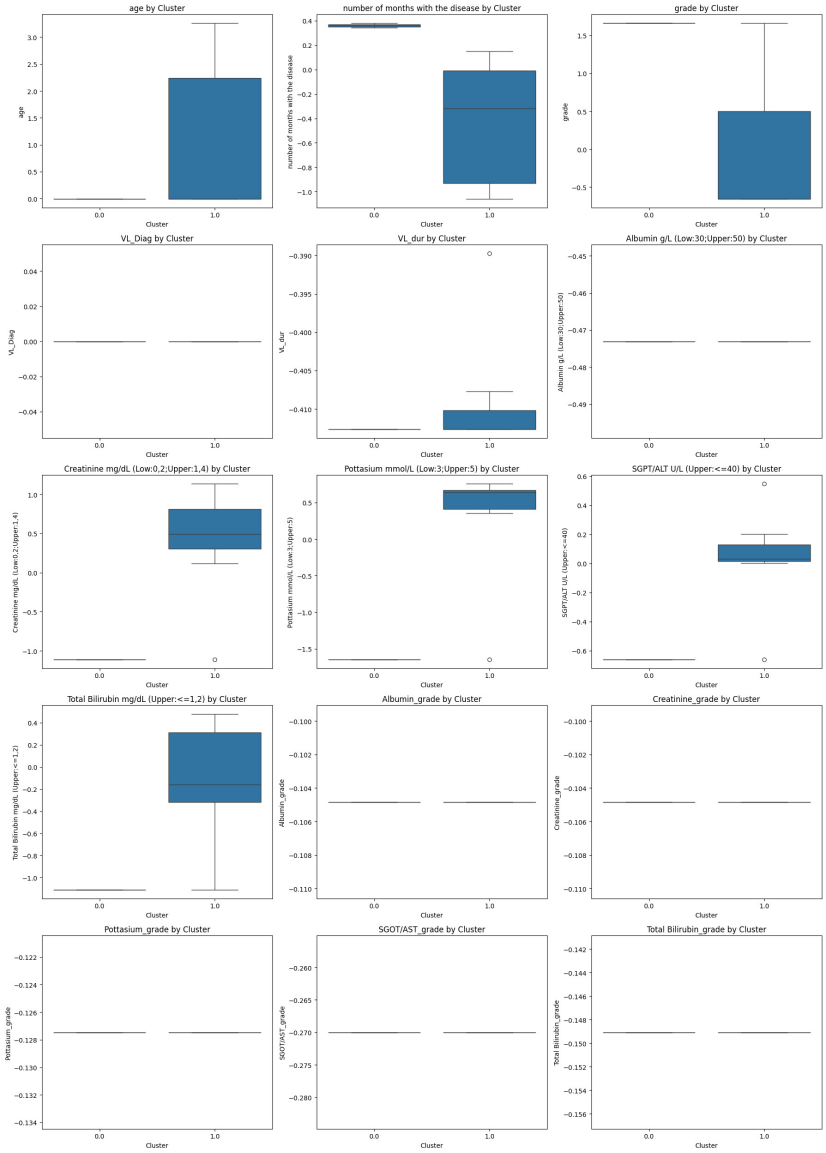
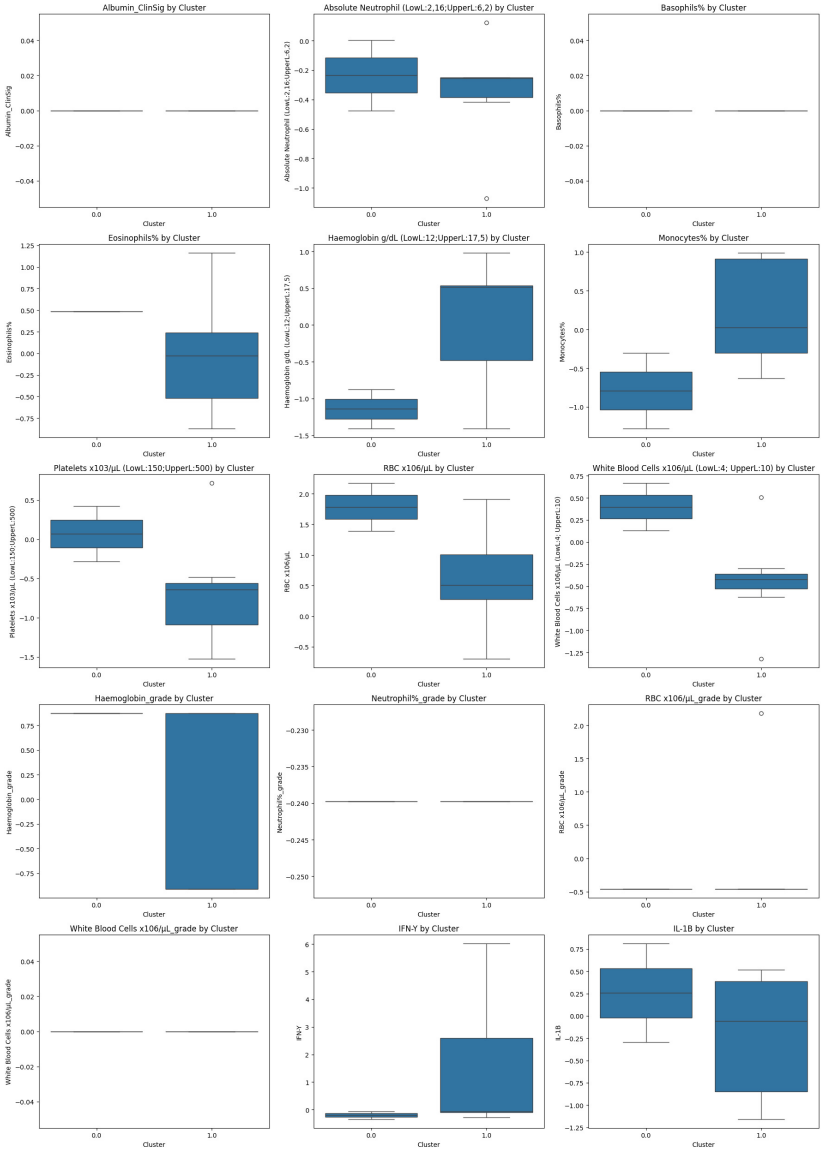


Figure 2: Boxplots in typical dataset of each variable across identified clusters. The plots summarize the distribution of values per cluster, highlighting central tendency, dispersion, and potential outliers. These visualizations provide insight into the discriminative power and internal variability of each feature within the clustering solution.

SUS Dataset


age by Cluster
number of months with the disease by Cluster
grade by Cluster
VL_Diag by Cluster
VL_dur by Cluster
Albumin g/L (Low:30;Upper:50) by Cluster
Creatinine mg/dL (Low:0,2;Upper:1,4) by Cluster
Pottasium mmol/L (Low:3;Upper:5) by Cluster
SGPT/ALT U/L (Upper:<=40) by Cluster
Total Bilirubin mg/dL (Upper:<=1,2) by Cluster
Albumin_grade by Cluster
Creatinine_grade by Cluster
Pottasium_grade by Cluster
SGOT/AST_grade by Cluster
Total Bilirubin_grade by Cluster
Cluster

Appendix


Albumin_ClinSig by Cluster
Absolute Neutrophil (LowL:2,16;UpperL:6,2) by Cluster
Basophils% by Cluster
Eosinophils% by Cluster
Haemoglobin g/dL (LowL:12;UpperL:17,5) by Cluster
Monocytes% by Cluster
Platelets x103/µL (LowL:150;UpperL:500) by Cluster
RBC x106/µL by Cluster
White Blood Cells x106/µL (LowL:4; UpperL:10) by Cluster
Haemoglobin_grade by Cluster
Neutrophil%_grade by Cluster
RBC x106/µL_grade by Cluster
White Blood Cells x106/µL_grade by Cluster
IFN-Y by Cluster
IL-1B by Cluster
Cluster

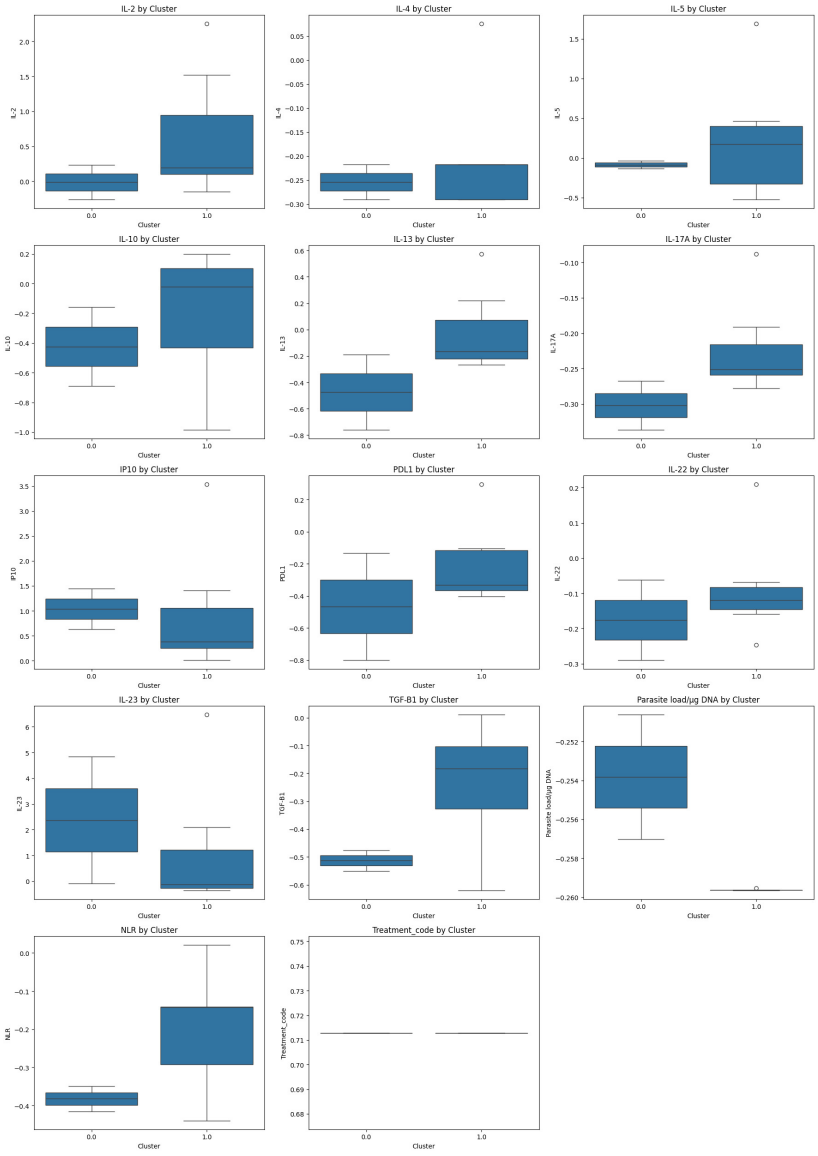
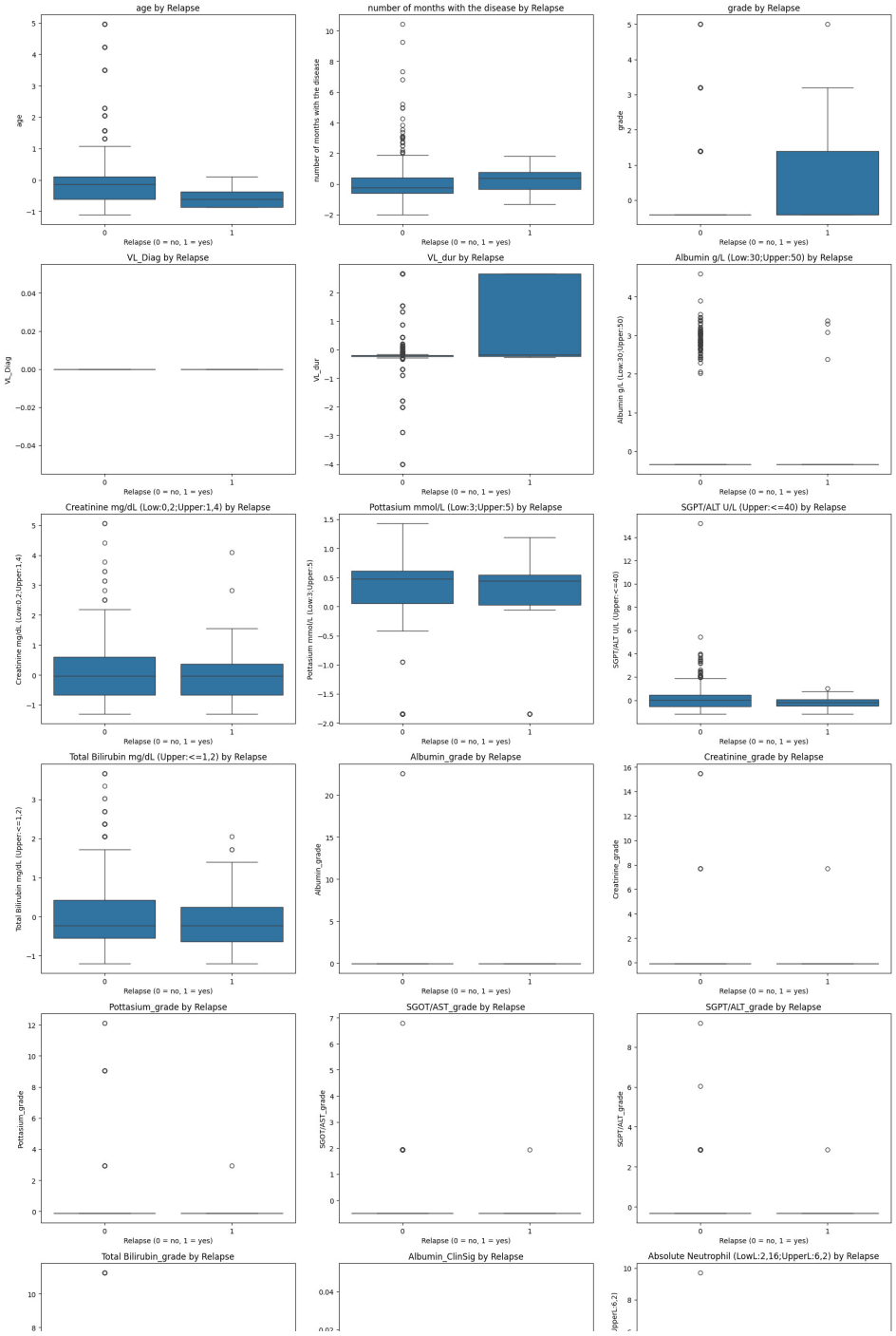


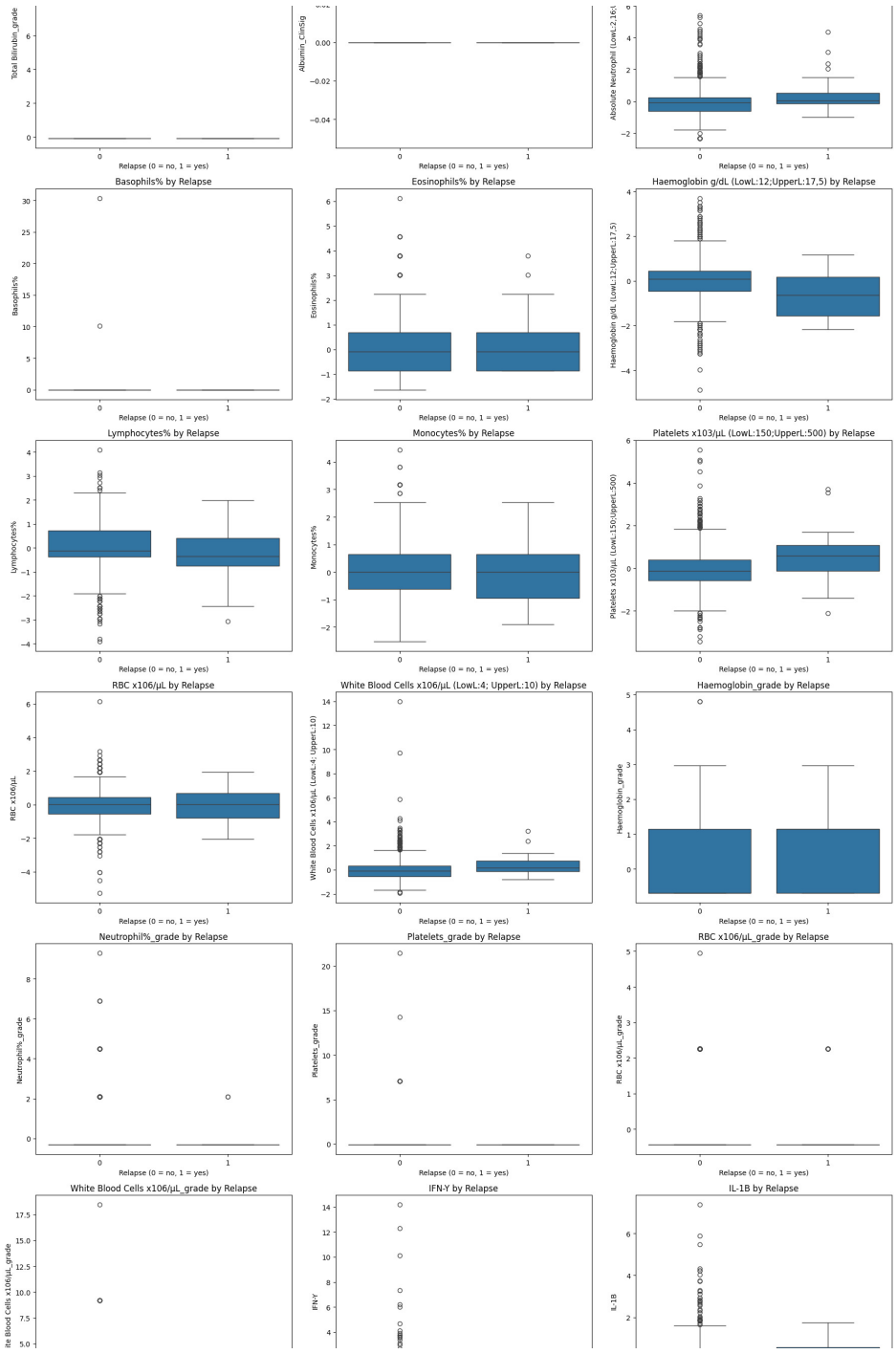
Figure 3: Boxplots in SUS dataset of each variable across identified clusters. The plots summarize the distribution of values per cluster, highlighting central tendency, dispersion, and potential outliers. These visualizations provide insight into the discriminative power and internal variability of each feature within the clustering solution.

.2. Relapse Boxplots

ML Dataset



Appendix


Total Bilirubin_grade
Albumin_ClinSig
Absolute Neutrophil (LowL:2,16)
Relapse (0 = no, 1 = yes)
Basophils% by Relapse
Basophils%
Eosinophils% by Relapse
Eosinophils%
Haemoglobin g/dL (LowL:12;UpperL:17,5) by Relapse
Haemoglobin g/dL (LowL:12;UpperL:17,5)
Lymphocytes% by Relapse
Lymphocytes%
Monocytes% by Relapse
Monocytes%
Platelets x103/µL (LowL:150;UpperL:500) by Relapse
Platelets x103/µL (LowL:150;UpperL:500)
RBC x106/µL by Relapse
RBC x106/µL
White Blood Cells x106/µL (LowL:4; UpperL:10) by Relapse
White Blood Cells x106/µL (LowL:4; UpperL:10)
Haemoglobin_grade by Relapse
Haemoglobin_grade
Neutrophil%_grade by Relapse
Neutrophil%_grade
Platelets_grade by Relapse
Platelets_grade
RBC x106/µL_grade by Relapse
RBC x106/µL_grade
White Blood Cells x106/µL_grade by Relapse
IFN-Y by Relapse
IFN-Y
IL-1B by Relapse
IL-1B

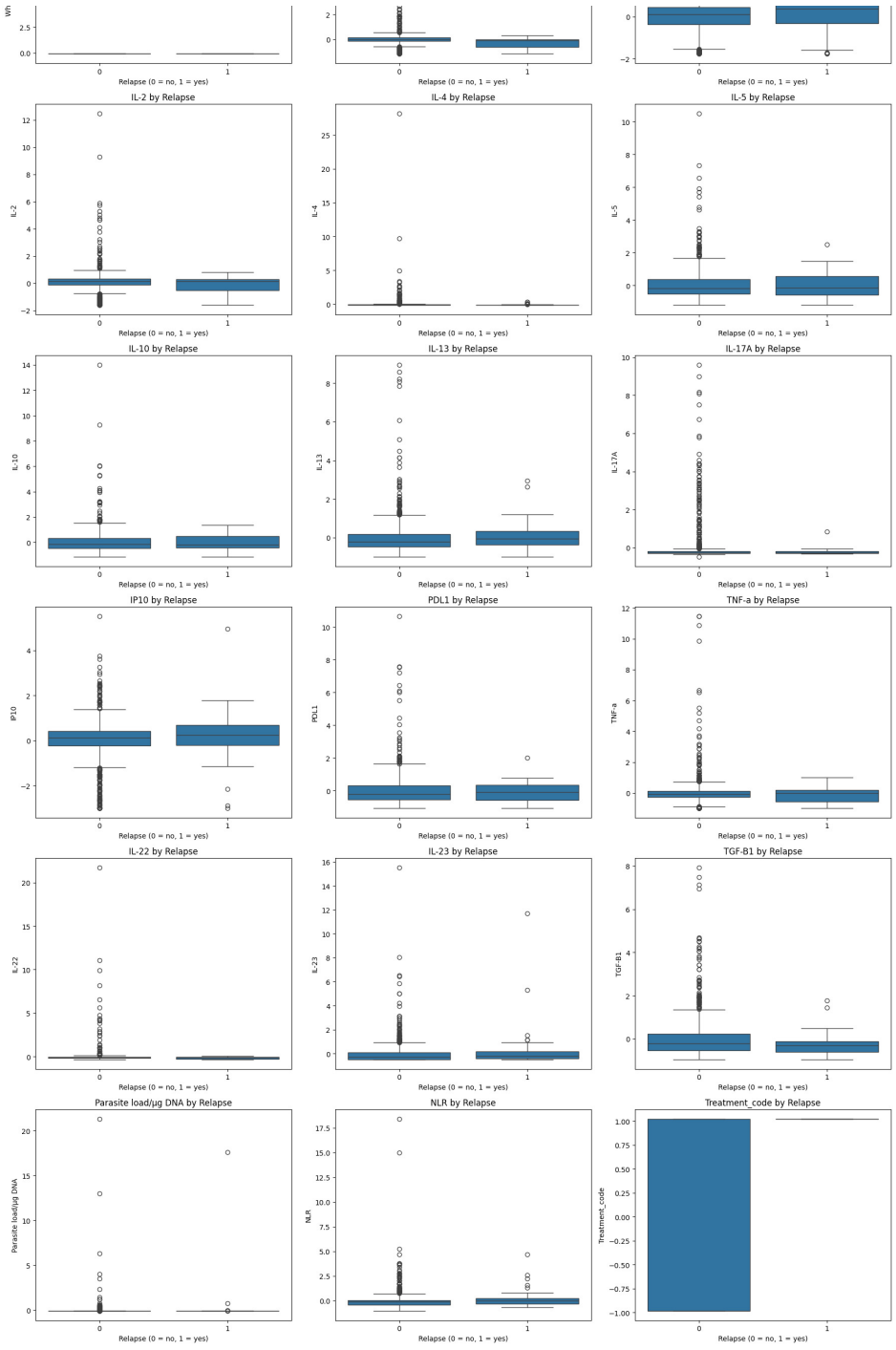
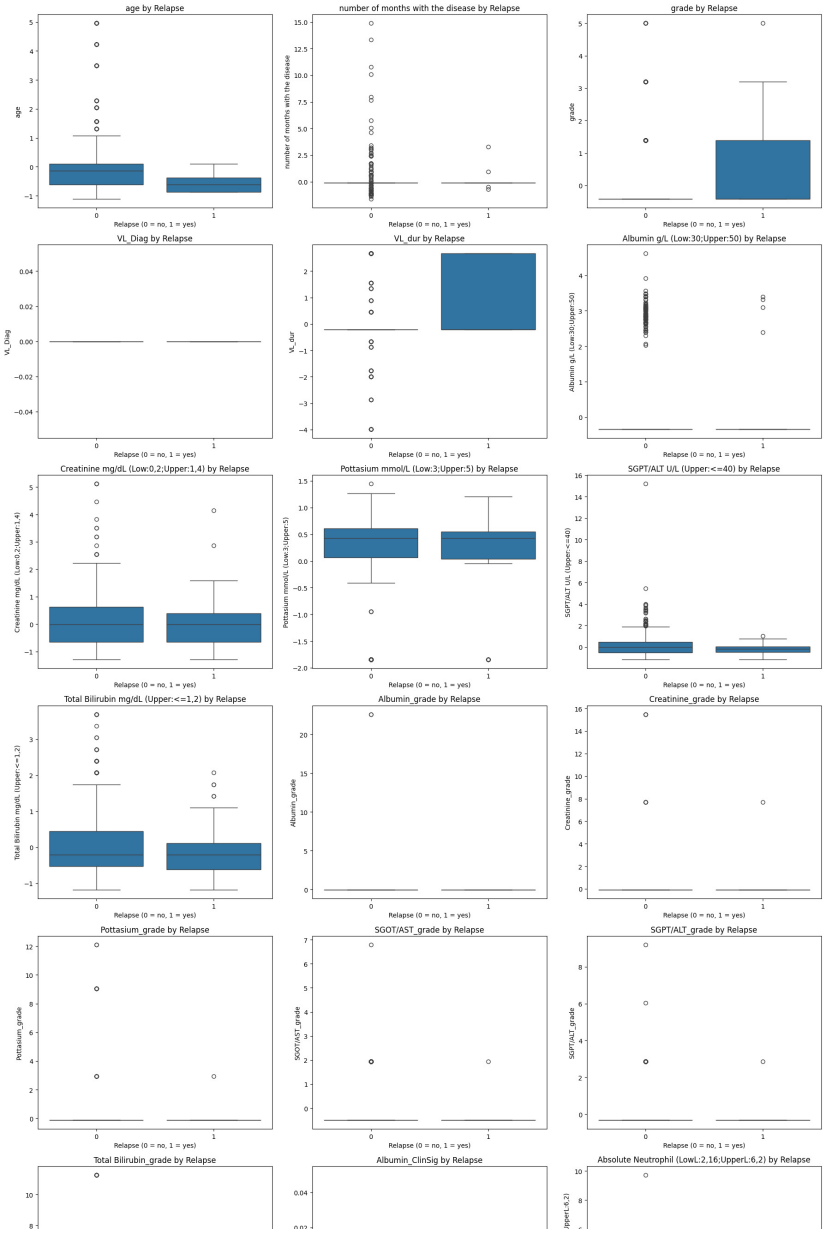
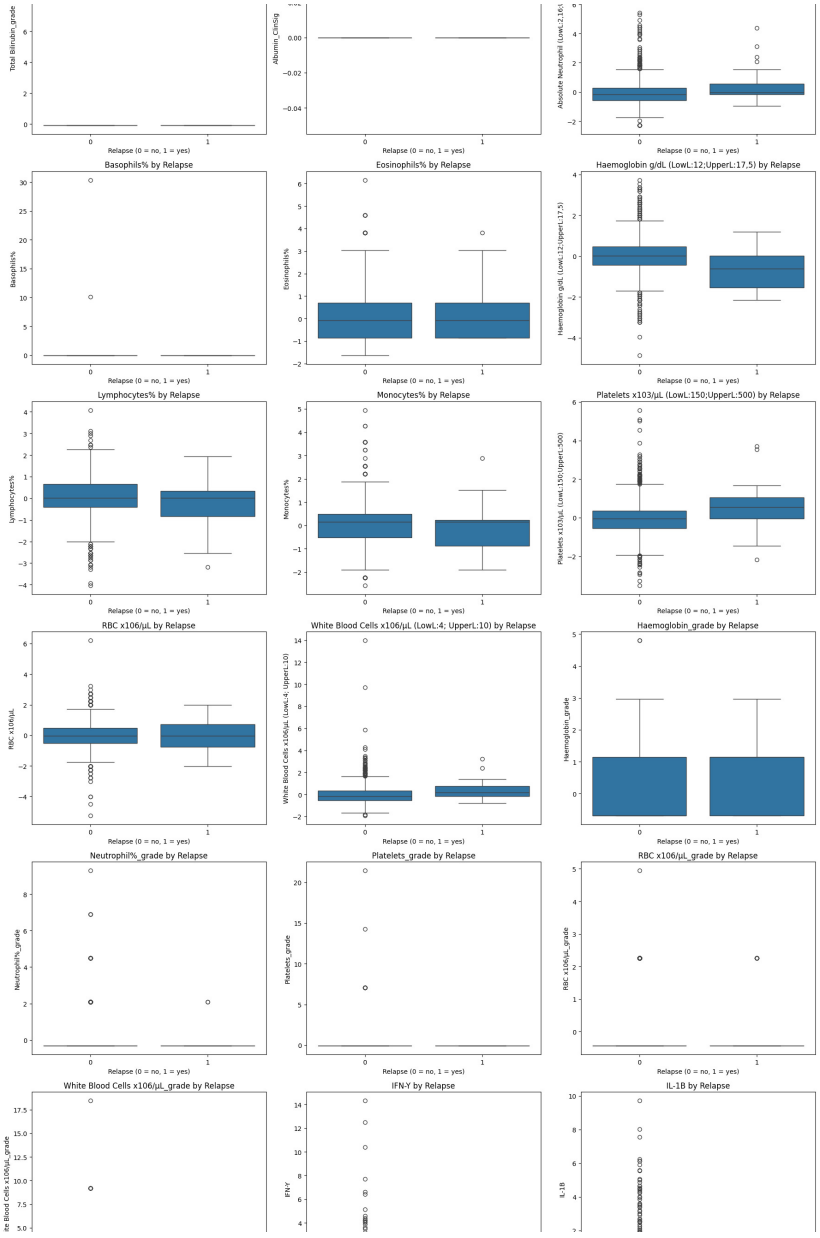


Figure 4: Boxplots in ML dataset of each variable across variable relapse. The plots summarize the distribution of values per relapse variable, highlighting central tendency, dispersion, and potential outliers. These visualizations provide insight into the discriminative power and internal variability of each feature within the clustering solution.

Typical Dataset




Basophils% by Relapse
Eosinophils% by Relapse
Haemoglobin g/dL (LowL:12;UpperL:17,5) by Relapse
Lymphocytes% by Relapse
Monocytes% by Relapse
Platelets x103/µL (LowL:150;UpperL:500) by Relapse
RBC x106/µL by Relapse
White Blood Cells x106/µL (LowL:4; UpperL:10) by Relapse
Haemoglobin_grade by Relapse
Neutrophil%_grade by Relapse
Platelets_grade by Relapse
RBC x106/µL_grade by Relapse
White Blood Cells x106/µL_grade by Relapse
IFN-Y by Relapse
IL-1B by Relapse
Relapse (0 = no, 1 = yes)

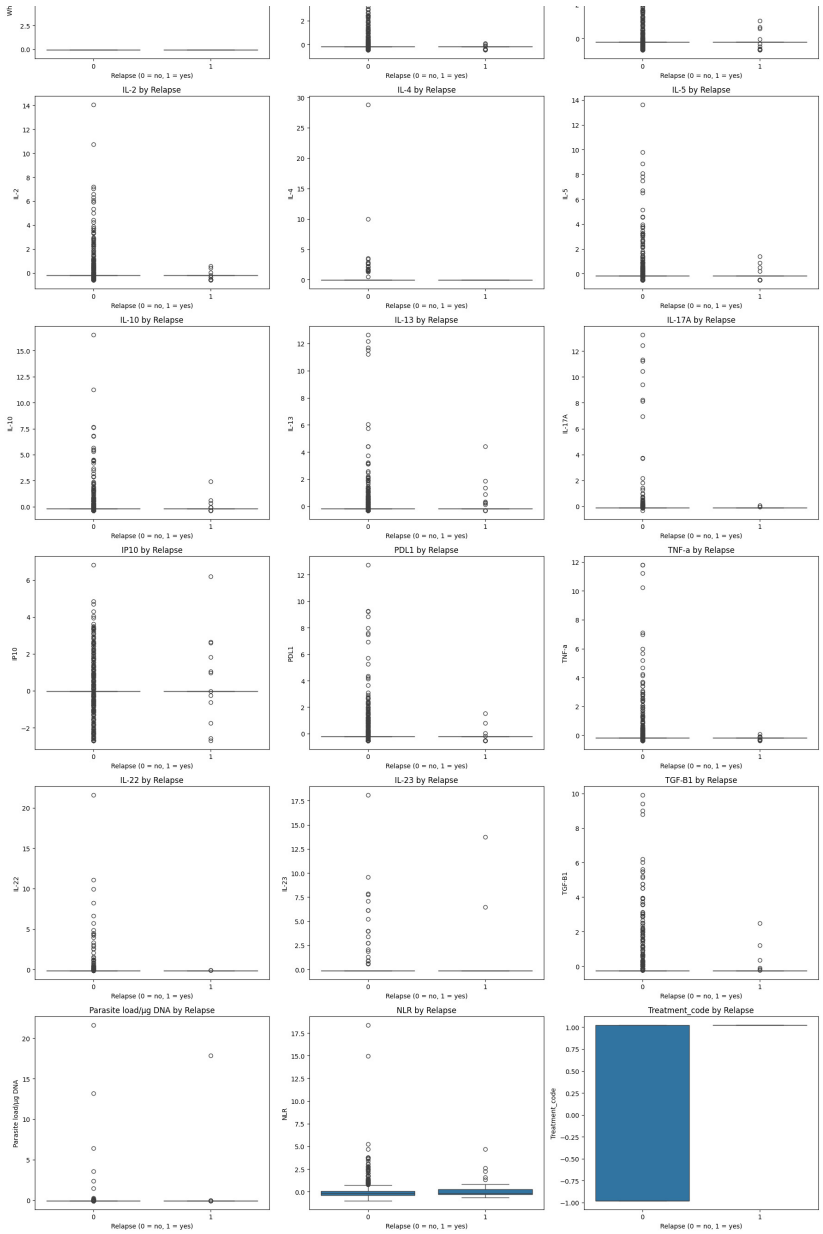
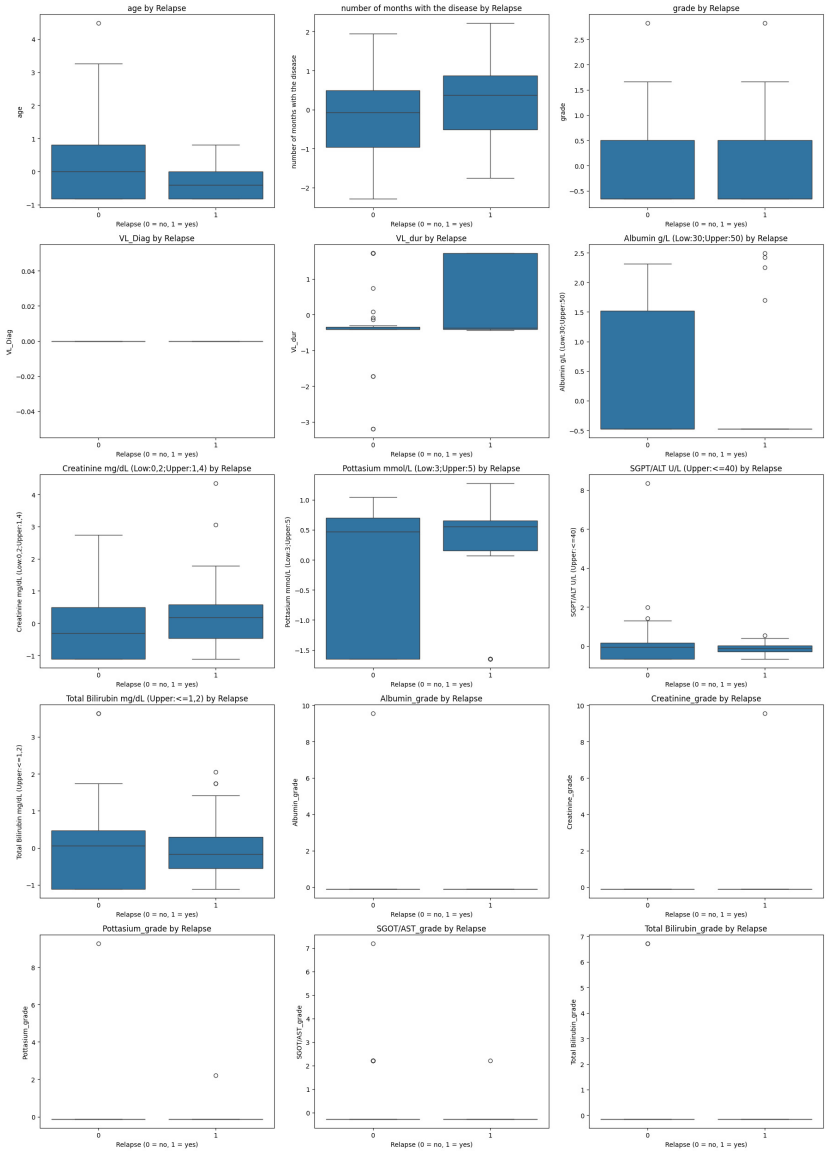
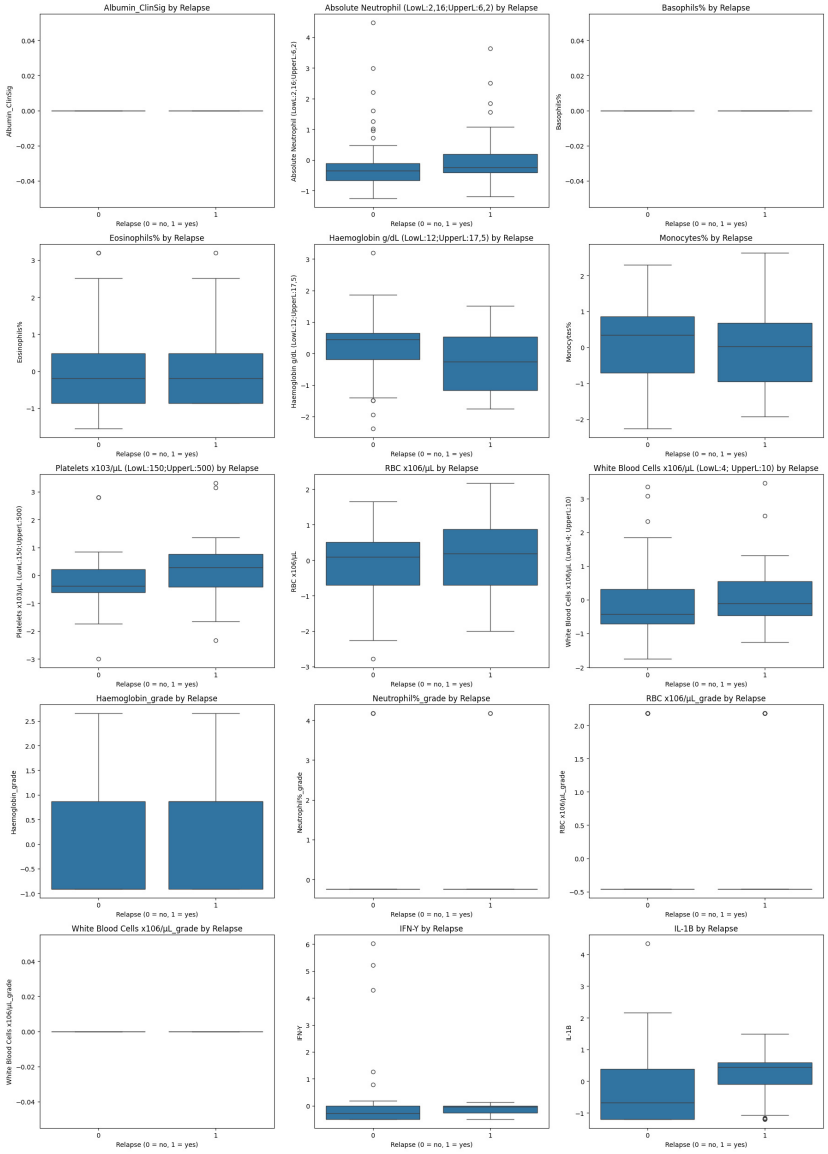


Figure 5: Boxplots in typical dataset of each variable across variable relapse. The plots summarize the distribution of values per relapse variable, highlighting central tendency, dispersion, and potential outliers. These visualizations provide insight into the discriminative power and internal variability of each feature within the clustering solution.

SUS Dataset


age by Relapse
number of months with the disease by Relapse
grade by Relapse
VL_Diag by Relapse
VL_dur by Relapse
Albumin g/L (Low:30;Upper:50) by Relapse
Creatinine mg/dL (Low:0,2;Upper:1,4) by Relapse
Pottasium mmol/L (Low:3;Upper:5) by Relapse
SGPT/ALT U/L (Upper:<=40) by Relapse
Total Bilirubin mg/dL (Upper:<=1,2) by Relapse
Albumin_grade by Relapse
Creatinine_grade by Relapse
Pottasium_grade by Relapse
SGOT/AST_grade by Relapse
Total Bilirubin_grade by Relapse
Relapse (0 = no, 1 = yes)

Appendix


Albumin_ClinSig by Relapse
Absolute Neutrophil (LowL:2,16;UpperL:6,2) by Relapse
Basophils% by Relapse
Eosinophils% by Relapse
Haemoglobin g/dL (LowL:12;UpperL:17,5) by Relapse
Monocytes% by Relapse
Platelets x103/µL (LowL:150;UpperL:500) by Relapse
RBC x106/µL by Relapse
White Blood Cells x106/µL (LowL:4; UpperL:10) by Relapse
Haemoglobin_grade by Relapse
Neutrophil%_grade by Relapse
RBC x106/µL_grade by Relapse
White Blood Cells x106/µL_grade by Relapse
IFN-Y by Relapse
IL-1B by Relapse
Relapse (0 = no, 1 = yes)

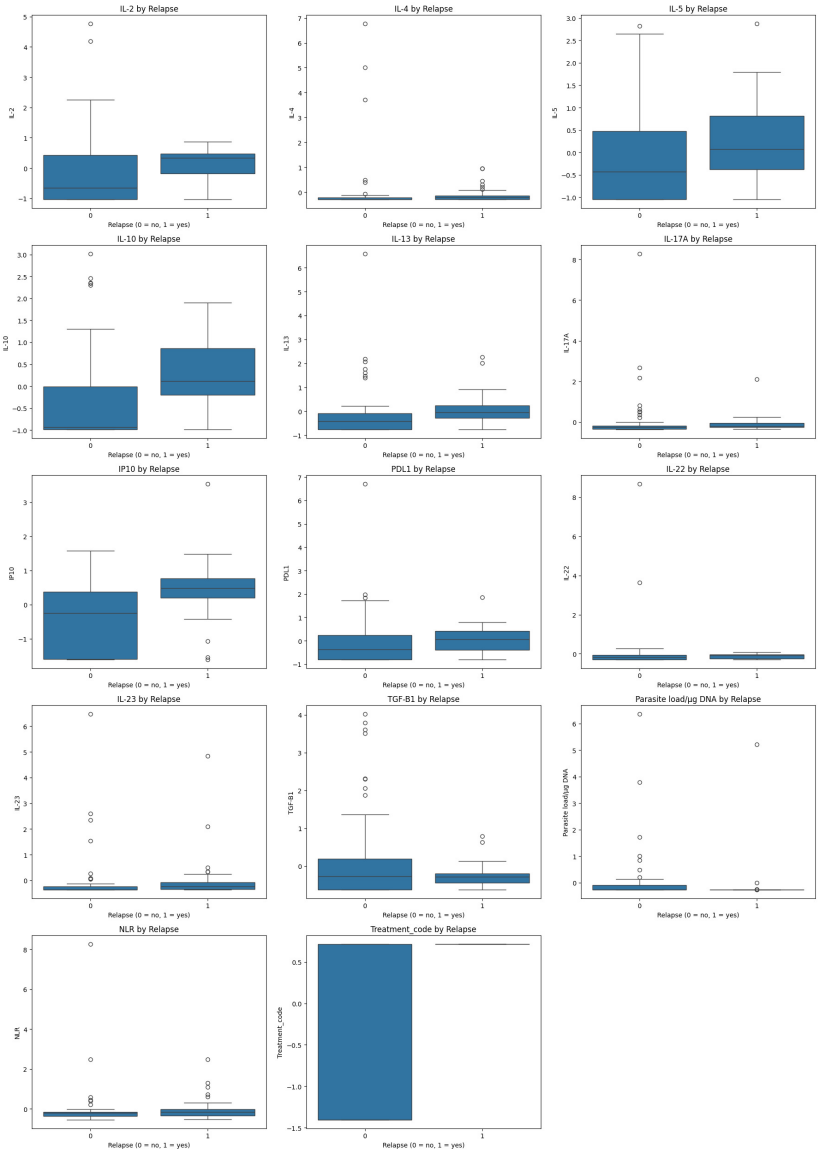


Figure 6: Boxplots in SUS dataset of each variable across variable relapse. The plots summarize the distribution of values per relapse variable, highlighting central tendency, dispersion, and potential outliers. These visualizations provide insight into the discriminative power and internal variability of each feature within the clustering solution.

Bibliography

[1] S. Ganguly, N. K. Das, J. N. Barbhuiya, and M. Chatterjee, "Post-kala-azar dermal leishmaniasis–an overview," *International journal of dermatology*, vol. 49, no. 8, pp. 921–931, 2010.

[2] D. Mukhopadhyay, J. E. Dalton, P. M. Kaye, and M. Chatterjee, "Post kala-azar dermal leishmaniasis: an unresolved mystery," *Trends in parasitology*, vol. 30, no. 2, pp. 65–74, 2014.

[3] V. Goyal, V. N. R. Das, S. N. Singh, R. S. Singh, K. Pandey, N. Verma, A. Hightower, S. Rijal, P. Das, J. Alvar *et al.*, "Long-term incidence of relapse and post-kala-azar dermal leishmaniasis after three different visceral leishmaniasis treatment regimens in bihar, india," *PLoS Neglected Tropical Diseases*, vol. 14, no. 7, p. e0008429, 2020.

[4] A. M. Musa, E. A. G. Khalil, B. Younis, M. Elfaki, M. Elamin, A. Adam, H. Mohamed, M. Dafalla, A. Abuzaid, and A. El-Hassan, "Treatment-based strategy for the management of post-kala-azar dermal leishmaniasis patients in the sudan," *Journal of tropical medicine*, vol. 2013, no. 1, p. 708391, 2013.

[5] A. Musa, E. Khalil, A. Hailu, J. Olobo, M. Balasegaram, R. Omollo, T. Edwards, J. Rashid, J. Mbui, B. Musa *et al.*, "Sodium stibogluconate (ssg) & paromomycin combination compared to ssg for visceral leishmaniasis in east africa: a randomised controlled trial," *PLoS neglected tropical diseases*, vol. 6, no. 6, p. e1674, 2012.

[6] A. Torres, B. M. Younis, M. Alamin, S. Tesema, L. Bernardo, J. C. Solana, J. Moreno, A.-a. Mustafa, F. Alves, A. M. Musa *et al.*, "Differences in the cellular immune response during and after treatment of sudanese patients with post-kala-azar dermal leishmaniasis, and possible implications for outcome," *Journal of Epidemiology and Global Health*, vol. 14, no. 3, pp. 1167–1179, 2024.

[7] E. E. Zijlstra, "The immunology of post-kala-azar dermal leishmaniasis (pkdl)," *Parasites & vectors*, vol. 9, no. 1, p. 464, 2016.

[8] C. Lopez, S. Tucker, T. Salameh, and C. Tucker, "An unsupervised machine learning method for discovering patient clusters based on genetic signatures," *Journal of biomedical informatics*, vol. 85, pp. 30–39, 2018.

[9] "Visual studio code - code editing. redefined." [Online]. Disponible: https://code.visualstudio.com/

[10] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay *et al.*, "Jupyter notebooks–a publishing format for reproducible computational workflows," in *Positioning and power in academic publishing: Players, agents and agendas*. IOS press, 2016, pp. 87–90.

[11] J. Hunter, "Matplotlib: A 2d graphics environment. computing in science &engineering 9, 90–95," 2007.

[12] M. L. Waskom, "Seaborn: statistical data visualization," *Journal of Open Source Software*, vol. 6, no. 60, p. 3021, 2021.

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[14] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[15] S. Seabold and J. Perktold, "Statsmodels: econometric and statistical modeling with python." *SciPy*, vol. 7, no. 1, pp. 92–96, 2010.

[16] J. Brownlee, *XGBoost With python: Gradient boosted trees with XGBoost and scikit-learn*. Machine Learning Mastery, 2016.

[17] C. M. Scavuzzo, J. M. Scavuzzo, M. N. Campero, M. Anegagrie, A. A. Aramendia, A. Benito, and V. Periago, "Feature importance: Opening a soil-transmitted helminth machine learning model via shap," *Infectious Disease Modelling*, vol. 7, no. 1, pp. 262–276, 2022.

[18] G. Varoquaux and O. Grisel, "Joblib: running python function as pipeline jobs," *packages. python. org/joblib*, 2009.

[19] B. M. Younis, A. Mudawi Musa, S. Monnerat, M. Abdelrahim Saeed, E. Awad Gasim Khalil, A. Elbashir Ahmed, M. Ahmed Ali, A. Noureldin,

G. Muthoni Ouattara, G. M. Nyakaya *et al.*, "Safety and efficacy of paromomycin/miltefosine/liposomal amphotericin b combinations for the treatment of post-kala-azar dermal leishmaniasis in sudan: A phase ii, open label, randomized, parallel arm study," *PLOS Neglected Tropical Diseases*, vol. 17, no. 11, p. e0011780, 2023.

[20] A. Torres, B. M. Younis, S. Tesema, J. C. Solana, J. Moreno, A. J. Martín-Galiano, A. M. Musa, F. Alves, and E. Carrillo, "Unsupervised machine learning identifies biomarkers of disease progression in post-kala-azar dermal leishmaniasis in sudan," *PLOS Neglected Tropical Diseases*, vol. 19, no. 3, p. e0012924, 2025.

[21] A. Gorshenin, M. Lebedeva, S. Lukina, and A. Yakovleva, "Application of machine learning algorithms to handle missing values in precipitation data," in *Distributed Computer and Communication Networks: 22nd International Conference, DCCN 2019, Moscow, Russia, September 23–27, 2019, Revised Selected Papers 22*. Springer, 2019, pp. 563–577.

[22] E. Acuna and C. Rodriguez, "The treatment of missing values and its effect on classifier accuracy," in *Classification, Clustering, and Data Mining Applications: Proceedings of the Meeting of the International Federation of Classification Societies (IFCS), Illinois Institute of Technology, Chicago, 15–18 July 2004*. Springer, 2004, pp. 639–647.

[23] J. Aleksic and M. García-Remesal, "A selective under-sampling (sus) method for imbalanced regression," *Journal of Artificial Intelligence Research*, vol. 82, pp. 111–136, 2025.

[24] P. Sedgwick, "Pearson's correlation coefficient," *Bmj*, vol. 345, 2012.

[25] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, "Unsupervised learning," in *An introduction to statistical learning: with applications in Python*. Springer, 2023, pp. 503–556.

[26] M. Fan, N. Gu, H. Qiao, and B. Zhang, "Intrinsic dimension estimation of data by principal component analysis," *arXiv preprint arXiv:1002.2050*, 2010.

[27] C. Garcia-Vidal, C. Teijón-Lumbreras, T. F. Aiello, M. Chumbita, R. Menendez, A. Mateu-Subirà, O. Peyrony, P. Monzó, C. Lopera, A. Gallardo-Pizarro *et al.*, "K-means clustering identifies diverse clinical phenotypes in covid-19 patients: implications for mortality risks and remdesivir impact," *Infectious Diseases and Therapy*, vol. 13, no. 4, pp. 715–726, 2024.

[28] V. Vardhan Baligodugula and F. Amsaad, "Unsupervised learning: Comparative analysis of clustering techniques on high-dimensional data," *arXiv e-prints*, pp. arXiv–2503, 2025.

[29] V. Kaverinskiy, I. Chaikovsky, A. Mnevets, T. Ryzhenko, M. Bocharov, and K. Malakhov, "Scalable clustering of complex ecg health data: Big data clustering analysis with umap and hdbscan," *Computation*, vol. 13, no. 6, p. 144, 2025.