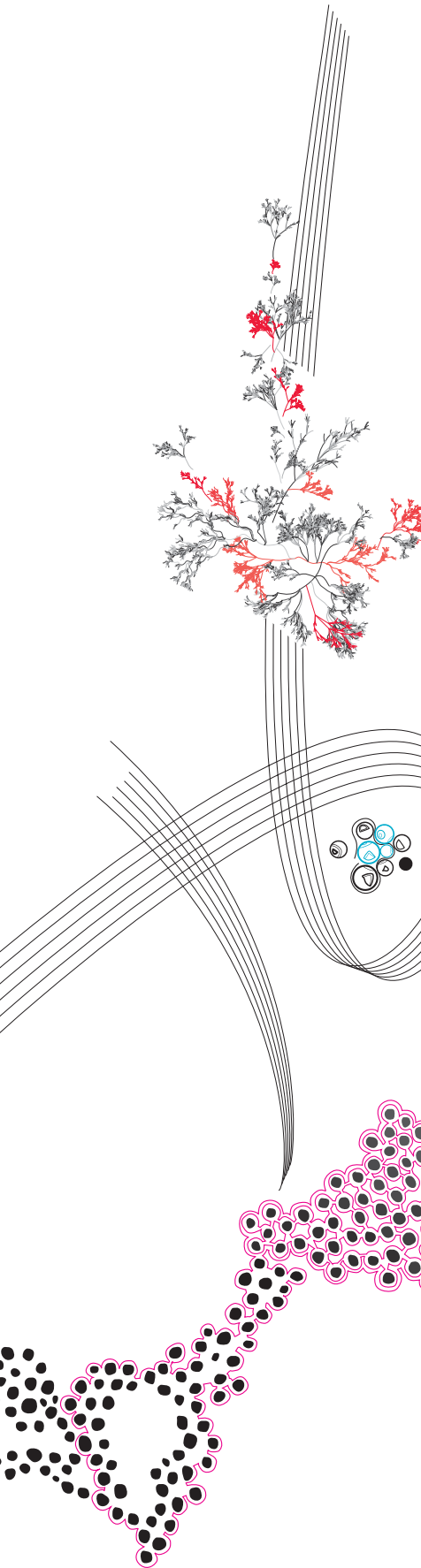


MSc Business Information Technology  
Final Project



# Enhancing Cancer Treatment Planning: A Combined Approach of Process Mining and Machine Learning with a Focus on Colorectal Cancer (CRC)

Student Lin Chin-Ying (Ally Lin)

Supervisor: Faiza.A.Bukhsh, Dr.A.Abhishta

March, 2024

**UNIVERSITY OF TWENTE.**

Department of Business Information Technology  
Faculty of Electrical Engineering,  
Mathematics and Computer Science,  
University of Twente

# Contents

<b>Chapter I</b>	<b>Introduction</b>	4
<b>Chapter II</b>	<b>Advanced Literature Review</b>	6
1	Related Work with Main Focuses	6
2	Search Strategy and Inclusion/Exclusion Criteria	7
3	Critical Appraisal of Collected Studies	7
4	Information Extraction and Synthesis Strategy	9
4.1	Tech Trends	10
4.2	Cancer Type	11
4.3	Medical Stage	11
5	Integration of Modern Technologies from Literature	13
6	Derivation of Research Gaps	13
<b>Chapter III</b>	<b>Design Science Methodology</b>	15
1	Engineering Cycle	15
2	Overview of the Method	16
<b>Chapter IV</b>	<b>Process Mining</b>	17
1	Data Collection and Preprocessing	17
2	Qualitative Analysis Using Process Mining	19
3	Summary	20
<b>Chapter V</b>	<b>Machine Learning</b>	21
1	Data Feature Extraction	21
2	Machine Learning Model Development and Training	24
3	Transfer Learning	25
4	Summary	26
<b>Chapter VI</b>	<b>Experimental Validation</b>	27
1	Explanation of the evaluation formula	27
2	Model Results	28
3	Summary	30
<b>Chapter VII</b>	<b>Discussion</b>	31
<b>Chapter VIII</b>	<b>Conclusion and Future Work</b>	33

# Abstract

Colorectal cancer (CRC) is a major health issue globally, highlighting the need for better treatment planning methods. Traditional decision-making processes in CRC treatment face obstacles due to the complex nature of individual patient cases, making it difficult to tailor treatment effectively. The combination of process mining and machine learning offers a fresh perspective for enhancing the accuracy of treatment decisions. This approach analyzes patient data to forecast the outcomes of different treatment options with greater precision.

The study uses clinical records from the MIMIC-III database, alongside sophisticated process mining tools (ProM), to find crucial information and patient features that impact CRC treatment results. It then trains machine learning models, particularly LSTM networks improved with transfer learning, to predict the survival rates of patients following various treatments. This method stands out for its use of process mining to map out treatment paths and machine learning to estimate the success of different treatments.

The result based on the specified evaluation method offers a confidence range of  $\pm 0.14$  (Avg\_EM), indicating that machine learning helps in analyzing and selecting the optimal treatment plans. This approach gives doctors a reliable range to customize patient care. Moreover, it helps patients to make informed choices about their treatment, reducing dependence on guesswork and uncertainty of the future.

In conclusion, this research not only demonstrates the practicality of merging process mining and machine learning to better CRC treatment planning but also paves the way for future studies, such as addressing data sparsity issues or further expanding this research direction. Therefore, this integrated approach could influence the future of personalized medicine and achieve the goal of establishing a data-driven healthcare system.

**Keywords:** Cancer Treatment Planning, Colorectal Cancer, Process Mining, Machine Learning, RNN, LSTM, Transfer Learning

# Chapter I

## Introduction

The field of cancer treatment planning is a critical and constantly developing aspect of medical science, characterized by its complex challenges. This paper focuses specifically on the complexities of colorectal cancer (CRC) treatment, a type of cancer notable for its high prevalence and intricacy. It is identified as the third most common and fourth deadliest cancer worldwide[38] and affects a broad spectrum of individuals across different genders and ages, from the elderly to younger people with increasingly aggressive forms[39]. This wide-ranging impact highlights the urgent need for refined and personalized treatment strategies.

Building upon this, the decision-making process in cancer treatment, a scenario frequently encountered in our lives, presents substantial challenges for patients and their families. This process, characterized by its complexity and the multitude of options available, demands careful consideration and often leads to significant stress and uncertainty for those affected. Navigating through many options and uncertainties, choosing a course of action becomes a formidable task. This challenge is particularly pronounced in CRC treatment, where choices range from invasive surgeries to targeted therapies[53], each with its own implications and varying success rates tailored to patients' diverse medical histories and individual conditions. Physicians often require many years of extensive experience to make informed judgments, yet providing a convincing rationale to patients facing decisions remains challenging due to the subjective nature of opinions. Thus, this underscores a significant gap in current treatment planning: the necessity for a data-driven, patient-specific approach that enables more precise and informed decision-making.

To meet this need, at the current stage of cancer treatment, significant progress has been made. For example, leveraging advancements in medical imaging, such as CT and MRI scans[47], and employing Natural Language Processing (NLP) to extract key information from Electronic Medical Records (EMRs)[22]. These technologies have significantly improved diagnostic accuracy. Furthermore, the increasing role of artificial intelligence, especially deep learning in medical image analysis, is crucial in enhancing treatment planning[9, 49]. Those papers also recognize the huge applicability of Machine Learning and Data Mining in predicting cancer outcomes, using techniques like biomarker models and sequence mining algorithms[29, 27]. Despite the apparent sophistication of current technologies, after drawing on a comprehensive review of advancements in colorectal cancer (CRC) diagnosis and treatment, the observations show that most of the papers still lack information on the application of process mining in CRC. As a result, it is crucial to **identify a solution that not only fills the existing gaps in current research but also leverages the predictive capabilities of machine learning to enhance the development of personalized strategies.**

This paper introduces a method that integrates Process Mining and Machine Learning to achieve a more objective approach. By jointly combining process mining with machine learning, it becomes feasible to develop more tailored treatment plans for individuals. To be more specific, Process Mining can be utilized to analyze detailed hospital event log data, offering a deep understanding of a patient’s treatment journey. This technique provides insights into the main treatment options, identifies patients’ key features, and finds a new way to support the facilitation of Transfer Learning. At the same time, Machine Learning can be applied to analyze a broad spectrum of health data, ranging from historical medical records to current symptoms, aiming to predict the outcomes of various treatment options and evaluate their post-treatment effects on patients’ quality of life. Therefore, this paper elaborates on the application of these methodologies in CRC treatment planning, striving to make cancer care more personalized and informed, thus reducing uncertainties and focusing on patient-centric personalized strategies.

The rest structure of the paper is outlined as follows: **Section II** presents a series of search strategies and literature reviews, clarifying the current use of similar technologies in cancer treatment and focusing on the origins and reasons for the research questions. **Section III** systematically summarizes the direction of the paper’s questions through the engineering cycle, taking both the artifact and the context into account to establish the clear objective and design problem, while also providing an overview of the entire methodology. **Section VI** and **Section V** detail the construction and application of process mining and machine learning, aimed at addressing the gaps, from data processing to the completed integration of technologies. **Section VI** establishes a specific formula to evaluate the model’s performance, considering the rationality of the results and their application.. **Section VII** discusses potential extended applications or current limitations and challenges of the technologies. Finally, **Section VIII** summarizes the essence of the paper, stating its contributions and importance, and discusses future research directions for feasibility. As a result, through this integrated approach, the paper aims to pave the way for a future where cancer treatment is not only more responsive but also deeply centered on the individual needs and conditions of each patient.

# Chapter II

## Advanced Literature Review

In the field of healthcare, numerous studies and applications related to process mining and machine learning exist. However, a consideration emerges in the context of family members facing dilemmas in medical scenarios. It questions whether the absence of anticipated answers or solutions in such challenging situations is attributable to the immaturity of available technology, or a lack of research on relevant topics at that time. In adherence to the guidelines[31, 12], our literature review adopts a comprehensive approach, establishing a structured framework for analyzing current cancer research, with a specific emphasis on colorectal cancer. This section outlines our research process, beginning with the formulation of the research's main focus, progressing through the search strategy, and continuing with the critical appraisal of the collected studies. Ultimately, a thorough examination of the selected literature will be conducted to identify unresolved issues at the current stage, serving as a means to clarify the starting point and reasons behind the primary discussion topics in this article.

### 1 Related Work with Main Focuses

The inquiry led to the formulation of three key Main Focuses (MFs). These focuses guide the critical analysis of recent literature related to cancer treatment planning and technological applications:

- **MF1: Treatment Planning for Colorectal Cancer**

Evaluate existing methods and techniques documented in recent literature on cancer diagnosis or therapy, specifically on colorectal cancer.

- **MF2: Empirical Effects and Applications of Process Mining and Machine Learning in Treatment Planning**

Examine notable empirical evidence and technological advancements associated with the integration of Process Mining and Machine Learning, particularly in medical treatment planning.

- **MF3: Identifying Gaps and Future Improvement Strategies in Colorectal Cancer Treatment**

Identify gaps or deficiencies existing in current colorectal cancer treatment planning and explore potential strategies for improvement, with a specific emphasis on the role of Process Mining and Machine Learning.

These Main Focuses (MFs) center on studies conducted over the last decade (2013 to 2023) across various disciplines, including Computer Science, Engineering, Health Professions, Decision Sciences, Immunology and Microbiology, Multidisciplinary, and Nursing. The investigation is particularly concentrated on colorectal cancer treatment strategies and the technological integrations employed in treatment planning.

## 2 Search Strategy and Inclusion/Exclusion Criteria

Following the Main Focuses (MFs), an extensive literature review was conducted using the reputable Scopus digital library, recognized for its extensive coverage of over 20,500 sources from 5,000+ publishers worldwide, providing researchers with a user-friendly, convenient, and comprehensive tool[19, 37, 12]. To align the retrieved literature with the main focuses of this paper, the key search query applied to the Article Title, Abstract, and Keywords of various papers was used:

*“(“colorectal cancer” OR “cancer”) AND (“treatment” OR “planning” OR “diagnosis” OR “therapy” OR “method” OR “technique” OR “approach”) AND (“process mining” OR “data mining”) AND (“machine learning” OR “artificial intelligence”)”*

The extensive screening process was thorough, focusing on articles published in the last ten years and within specific relevant subject areas. Duplicate articles were also identified and removed from this selection. Moreover, to ensure a thorough yet targeted collection of relevant studies, inclusion/exclusion criteria were implemented. The detailed criteria are as follows:

### **Inclusion Criteria (ICs):**

- **IC1:** Addresses the formulated Main Focuses.
- **IC2:** Document Type is Article, Conference Paper, or Review.
- **IC3:** Available in English Journal and accessible for download.

### **Exclusion Criteria (ECs):**

- **EC1:** The main topic and contents of the paper are not significantly related to the research focus of this article.
- **EC2:** The paper does not mention specific methods to solve problems.

The primary focus was on articles from the categories of Article, Conference Paper, and Review. These categories were broadly classified into detailed and experiment-specific descriptions or comprehensive systematic introductions to a single field. By filtering these two main categories, a macroscopic reading of Review content was possible, followed by a more in-depth and microscopic understanding of technical details and methods from Articles and Conference Papers. Additionally, only articles published in English-language Journals and available for download were selected. This was done to ensure and facilitate a more in-depth examination of the articles later. Lastly, to ensure relevance to the article’s main topic, articles less related to the subject were filtered out. This strategy guarantees a comprehensive collection of pertinent studies. Therefore, through these screening strategies and initial readings, a more in-depth study can be conducted on the remaining articles.

## 3 Critical Appraisal of Collected Studies

In the critical appraisal phase, the collected studies undergo a rigorous evaluation. This process aims to determine their contribution to the understanding of colorectal cancer treatment planning, especially in the realm of process mining and machine learning. The appraisal of the studies is based on several key criteria [FIGURE 1]:

<p><b>(1) Relevance to Colorectal Cancer Treatment Planning</b> [Scoring: 0 - No mention of colorectal cancer 1 - Technology mentioned with validation using colorectal or multiple cancer data 2 - Using colorectal cancer and one or two other cancers as primary research subjects 3 - Colorectal cancer as the main research goal] * As the focus of this paper is mainly on the potential of technology with colorectal cancer as the primary application target, rather than an in-depth study of colorectal cancer itself, the maximum score for this criterion is set at 3, to not overlook articles with relevant technological value.</p> <p><b>(2) Impact on Medical Decision-Making Stage</b> [Scoring: 0 - General mention of decision-making stages without specific insights 1 - Early prediction and treatment 3 - Post-treatment side effects and follow-ups 5 - Strong relevance to treatment choices and their impact on survival rates]</p> <p><b>(3) Use of Process Mining Techniques</b> [Scoring: 0 - Low relevance to the technology application of this paper 1 - Mentioned but not elaborated or utilized effectively 2 - Some relevance with basic use or examples 3 - Content can serve as a reference 4 - Well-integrated use of process mining techniques 5 - Strong relevance to the technology direction of this paper]</p> <p><b>(4) Use of Machine Learning Techniques</b> [Scoring: 0 - Low relevance to the technology application of this paper 1 - Mentioned but not elaborated or utilized effectively 2 - Some relevance with basic use or examples 3 - Content can serve as a reference 4 - Well-integrated use of machine learning techniques 5 - Strong relevance to the technology direction of this paper]</p>	<p><b>(5) Clarity and Methodological Soundness</b> [Scoring: 0 - Primarily literature review 1 - Basic methodology described without details 2 - Clear methodology but lacks depth or rigor 3 - Good methodological description with some application details 4 - Very clear and detailed methodological description 5 - Detailed explanation of methods used and their application]</p> <p><b>(6) Empirical Validation</b> [Scoring: 0 - No mention 1 - Mentioned but lacks detail or relevance 2 - Some empirical evidence but limited in scope or depth 3 - Good empirical validation with relevant results 4 - Strong empirical validation with comprehensive data 5 - Detailed description and validation process]</p> <p><b>(7) Innovation and Future Potential</b> [Scoring: 0 - No mention 1 - Mentioned but generic or speculative 2 - Some innovative aspects with potential implications 3 - Notable innovation and relevant to current and future research 4 - Significant innovation with strong future potential 5 - Content strongly relevant to this paper and can aid in further extended discussions]</p>
---	---

FIGURE 1: Detailed Scoring Criteria

**(1) Relevance to Colorectal Cancer Treatment Planning (w.r.t MF1)**

Studies are evaluated based on their relevance to colorectal cancer. Although technologies applicable in various domains may offer relevant insights and serve as references for subsequent experiments, studies specifically addressing colorectal cancer are particularly valuable for acquiring pertinent information and delving into related research areas. Therefore, studies offering significant insights into colorectal cancer, either through empirical data or theoretical propositions, receive higher scores.

**(2) Impact on Medical Decision-Making Stage (w.r.t MF1)**

The primary focus of this criterion is on the stage following patient diagnosis, emphasizing the identification of tools that assist in decision-making between patients and doctors. This differs from most articles that concentrate on diagnostic prediction and early treatment. Instead, this criterion seeks to identify studies pertinent to prognosis, which may include aspects such as survival rate predictions, recurrence warnings, or recommendations for medication and diagnostic follow-ups. Consequently, articles that contribute to medical decision-making, especially those that provide insights into post-diagnosis stages, are awarded higher scores.

**(3) Use of Process Mining Techniques (w.r.t MF2)**

The depth and extent of process mining utilization in each study are examined. The evaluation criteria favor studies that show a strong connection to the technologies researched in this paper. This is because most articles primarily concentrate on data mining, aiming to uncover key molecular mechanisms in various conditions. As such, studies that mention process mining and integrate it effectively into their research methodology will score higher.



(4) **Use of Machine Learning Techniques (w.r.t MF2)**

The depth and complexity of machine learning utilization in each study are examined. This includes the methodologies applied, the data sources used, and the results achieved. As discussions span Artificial Intelligence or Deep Learning, focusing on Machine Learning algorithms will aid in selecting models and parameters. Thus, articles with strong relevance to the technologies and applications studied in this paper will score higher.

(5) **Clarity and Methodological Soundness (w.r.t MF2)**

The emphasis is on the clarity and methodological rigor of the studies. Those that are methodologically sound and present their findings and methodologies clearly, in a replicable and well-documented manner, score higher.

(6) **Empirical Validation (w.r.t MF3)**

This criterion measures the extent of empirical validation provided by the studies, involving looking into the research methodologies, data robustness, and credibility of the results, with an exploration of their validation stages' completeness and effectiveness.

(7) **Innovation and Future Potential (w.r.t MF3)**

This criterion values studies that address current deficiencies in cancer treatment planning and propose novel approaches or strategies for the future. It includes the exploration of emerging technologies such as XAI (Explainable Artificial Intelligence) and NLP (Natural Language Processing), as well as considerations for future technological predictions, and aspects related to the protection and storage of medical data privacy. These elements can aid in refining methods, offering more comprehensive technological insights, and fostering extended discussions in future research endeavors.

A scoring system is used for systematic appraisal, assisting in filtering the literature more helpful to this paper. Studies scoring below a certain degree are considered less relevant to the analysis and are excluded. This threshold ensures the inclusion of only high-quality, relevant studies in the analysis. Therefore, the critical appraisal phase plays a pivotal role in refining the literature base. It ensures reliance on high-quality studies that are directly relevant to the paper's main focus and questions, offering substantial insights into the complexities of colorectal cancer treatment planning with process mining and machine learning. The findings from these appraised studies lay the foundation for subsequent sections of the paper, aiming to establish a robust research process and make significant contributions to colorectal cancer treatment.

## 4 Information Extraction and Synthesis Strategy

Building on the earlier sections, this section forms a crucial step in our comprehensive review of the literature on cancer treatment planning. This phase is dedicated to synthesizing the results obtained from the previously outlined steps – concentrating on the main research focuses, applying the search strategy with inclusion and exclusion criteria, and conducting a critical appraisal of the collected studies. The objective is to extract valuable insights, identify trends, and consolidate findings to advance the understanding of cancer treatment planning, particularly in the context of Colorectal Cancer, Process Mining, and Machine Learning.

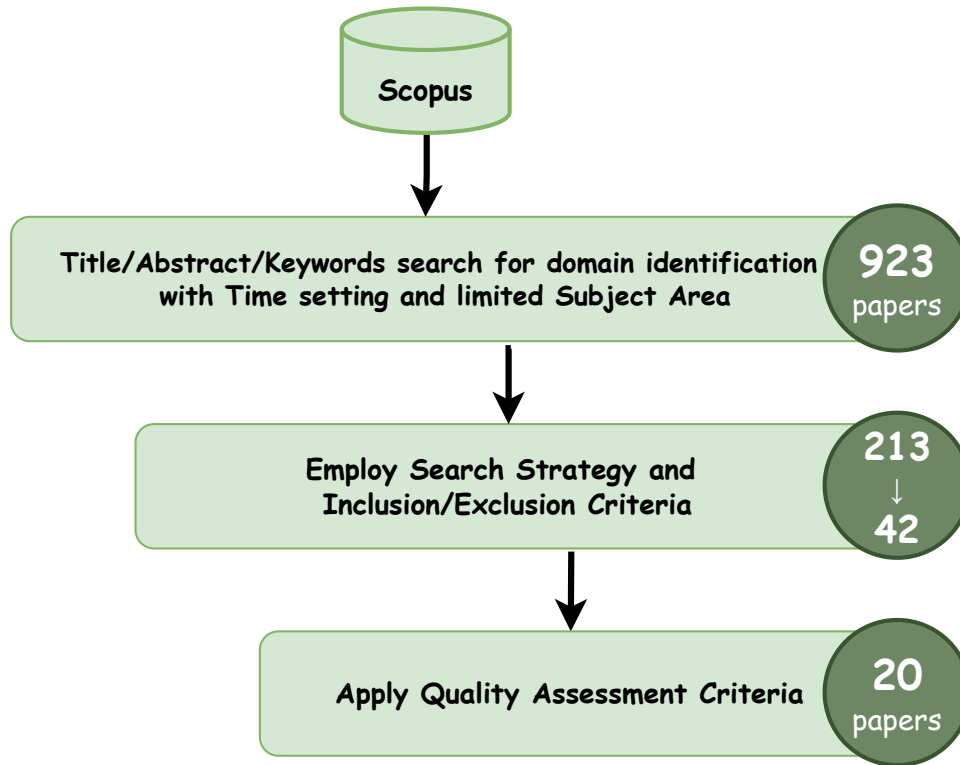
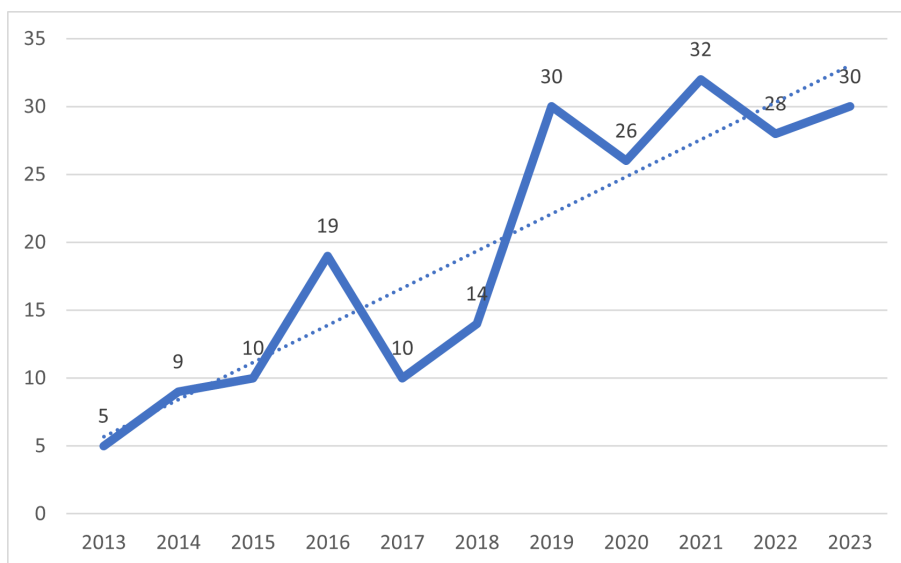


FIGURE 2: Article Selection Process and Results

#### 4.1 Tech Trends

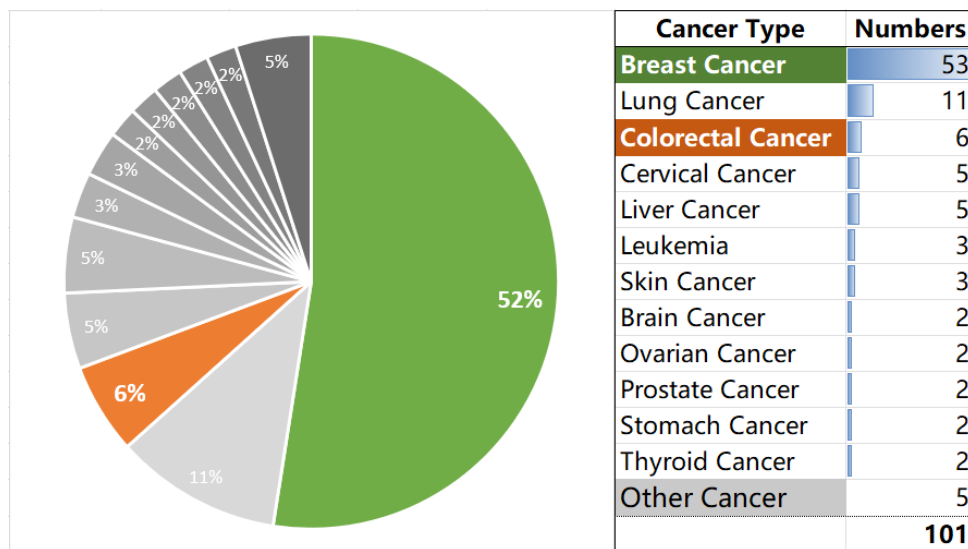
Initially, an extensive search yielded 923 articles, selected for their relevance to the main focuses, based on their titles, abstracts, and keywords. This initial pool was refined down to 213 articles using specific search strategies with inclusion criteria [FIGURE 2], ensuring that the selected studies directly addressed the main research areas. Meanwhile, the analysis of the collected articles revealed several noteworthy trends, particularly in the realm of technology application within cancer research. One particularly striking trend was the recent and marked increase in the use of process mining and machine learning technologies [GRAPH 1]. This upswing highlights a growing interest among scholars in these advanced techniques. The trend suggests a broader acknowledgment and adoption of these technologies in recent years, as evidenced by the growing number of publications in this area.



GRAPH 1: Articles by Year

## 4.2 Cancer Type

Despite these technological advancements, an apparent gap was identified in the type of cancer predominantly researched. Among 213 articles analyzed, 112 primarily discussed technology without focusing on any specific type of cancer, classified under the 'Unspecified' category. However, among the remaining 101 articles, most of the existing research is centered around breast cancer [GRAPH 2]. This trend, while indicative of the considerable attention given to breast cancer, simultaneously underscores a deficiency in research specifically directed toward Colorectal Cancer (CRC). This disparity has shown the importance of focusing more on CRC, recognizing its relatively unexplored potential in research. By centering the investigation on CRC, the aim is to generate robust data that transcends the traditional gender-specific focus of cancer research. Such an approach not only enriches the understanding of CRC but also expands the reach and relevance of the findings, making them more universally applicable and inclusive.



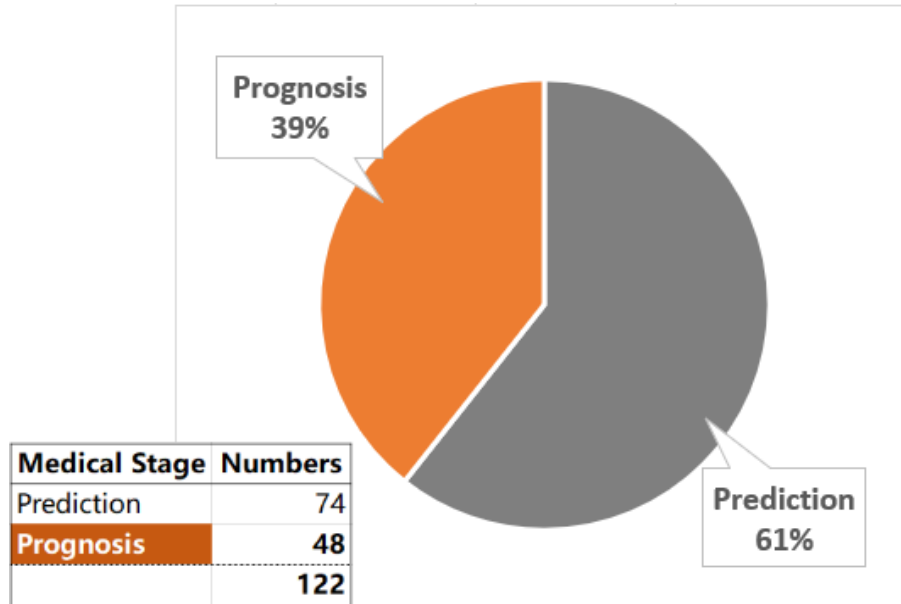
GRAPH 2: Frequency of Each Cancer Type

## 4.3 Medical Stage

Moreover, an observation was noted regarding the stages of cancer that most studies tend to address. The studies were categorized into three distinct areas:

- Prediction Stage:** This phase occurs when the patient's health status is unknown. Techniques such as data mining and machine learning are employed, along with methods like genetic profiling, imaging comparisons, and classification. These approaches are used to predict potential risks or future health developments, aiding in early cancer prevention and identifying latent health risks.
- Prognosis Stage:** This stage begins after a patient has been diagnosed with an illness. It involves a detailed assessment of the disease to understand its potential progression and severity. This stage is crucial for patients and healthcare teams to make informed treatment decisions, offering predictive recommendations to tackle future challenges. The prognosis stage focuses primarily on evaluating the disease's progression and formulating treatment plans post-diagnosis.
- Comprehensive Analysis Stage:** This stage involves research that benefits multiple aspects of a patient's medical condition, often without a clear boundary. Research in this stage provides insights that cover the entire healthcare continuum, offering a holistic view of factors influencing patient health, treatment strategies, and potential future advancements in medical care.

Significantly, excluding 91 articles in the comprehensive analysis stage, as shown in the comparison between prediction and prognosis, the bulk of research focuses on the early stages of cancer, centering on prevention and initial diagnosis [GRAPH 3]. Nevertheless, there seems to be less emphasis on the subsequent medical processes for those who have already been diagnosed with cancer, including treatment selection, prognostic evaluation, and predictions of post-treatment survival rates. Thus, this observation suggests that this area of research holds great potential and could significantly contribute to the planning of cancer treatment.



GRAPH 3: Frequency of Two Main Medical Stages

Subsequently, further refinement using exclusion criteria resulted in a narrowed selection of 42 highly relevant articles, [FIGURE 2]. These articles underwent a rigorous quality assessment process, from which only papers scoring above 15 points were chosen for the final analysis [Appendix A]. These papers were evaluated based on their relevance, impact on medical decision-making, use of process mining and machine learning techniques, clarity, methodological soundness, empirical validation, and innovation potential. These stringent criteria ensured that only the most pertinent and high-quality papers were included in the final analysis phase.

After filtering out 20 articles that met the criteria, the final step entailed a comprehensive analysis of the selected papers to derive key insights, trends, and patterns relevant to supporting this research topic. Utilizing the previously established detailed scoring criteria, it became clear to identify the value of each article based on its high-scoring aspects. These include detailed explanations about colorectal cancer or recommendations regarding the selection of machine learning models. The information gathered from these analyses was then deliberated upon in terms of its impact on cancer treatment planning. This discussion involved assessing the potential of process mining and machine learning to enhance treatment strategies, pinpointing areas for future research, and contemplating the practical implementation of these technologies in clinical settings. Overall, this rigorous process of data extraction and synthesis aims to significantly deduce and support the main topic and research questions of this article. By integrating insights from process mining and machine learning research, the study offers a comprehensive understanding of the current challenges and prospective advancements in CRC treatment, thus further supporting the contribution and significance of this article.

## 5 Integration of Modern Technologies from Literature

Based on the comprehensive review of the latest advancements in cancer treatment, particularly colorectal cancer (CRC), significant progress has been made in both diagnosis and therapy. Advanced medical imaging techniques like Computer Tomography (CT) and Magnetic Resonance Imaging (MRI) have revolutionized CRC diagnosis, offering detailed insights into the tumor's location and stage[47]. Moreover, Electronic Medical Records (EMRs), especially unstructured consultation notes, have emerged as a crucial information source[22]. Utilizing Natural Language Processing (NLP) technologies to extract key predictive factors for CRC from these notes has become increasingly significant[22]. In terms of treatment, a combination of surgery, chemotherapy, and radiation therapy is typically employed, varying according to the cancer stage and the patient's overall health[53]. The role of artificial intelligence and machine learning in diagnosing and treating CRC is also growing in importance. For instance, deep learning models are now being used in medical image analysis to identify tumor areas more accurately[9, 49].

At present, Machine Learning and Data Mining technologies show significant potential in the field of cancer treatment research. These technologies, particularly multi-stage learning methods, biomarker models, and sequence mining algorithms, have been applied in predicting the survival and recurrence risks of breast cancer, cervical cancer, acute myeloid leukemia, ovarian cancer, and prostate cancer[35, 34, 25, 29, 27]. These studies underscore the importance of extracting key features from clinical data to improve the accuracy of prediction models[25]. For example, research on breast cancer recurrence has demonstrated that combining statistical feature selection, multi-classifier evaluation, and Brainstorm Optimization (BSO) can effectively enhance the performance of prediction models[6]. Similarly, in the prognosis assessment of cervical cancer, models based on the DNA methylation of four CpG sites have shown a high predictive capability[35]. Additionally, the use of Convolutional Neural Networks (CNN) in processing high-dimensional gene expression data, especially under transfer learning strategies, has shown superior performance in predicting lung cancer survival, offering higher predictive accuracy and sensitivity compared to traditional machine learning methods[34]. However, there's a notable deficiency in the application of process mining within the cancer treatment domain, specifically in CRC. Process mining, a tool focused on extracting knowledge from event logs to understand and optimize actual processes, can significantly enhance the treatment journey for CRC patients. Similar technologies include the use of sequence mining algorithms and classification methods, considering treatment time intervals and patient quality of life, to predict patient survival outcomes[29, 27]. Additionally, process mining can help identify bottlenecks in treatment processes and provide more personalized treatment recommendations, thereby enhancing patients' life quality and treatment efficacy.

## 6 Derivation of Research Gaps

After a thorough examination and understanding of the literature related to cancer treatment, these can response to the initial main focuses to uncover existing gaps:

- **Response to the MF1: Treatment Planning for Colorectal Cancer**

Despite significant advancements in diagnostic and treatment methods, including the application of machine learning in medical imaging analysis and prediction models, there is a pressing need for continual improvement. Enhancing the accuracy of tumor detection through comprehensive clinical data analysis and improving the predictive accuracy of treatment outcomes remain crucial areas for future research.

- **Response to the MF2: Empirical Effects and Applications of Process Mining and Machine Learning in Treatment Planning**

The literature indicates a noticeable deficiency in the application of process mining within the CRC treatment domain. The potential of process mining to extract knowledge from event logs, visualize critical information, and construct overall treatment pathways has not been fully utilized. Expanding the application of process mining and further integrating it with machine learning to predict patient outcomes and improve treatment efficacy is an important research direction.

- **Response to the MF3: Identifying Gaps and Future Improvement Strategies in Colorectal Cancer Treatment**

Although progress has been made in machine learning and data mining individually for cancer treatment, the specific application of these technologies in CRC, especially in personalized treatment planning, has been overlooked. Future research should focus on more effectively integrating process mining and machine learning, concentrating on integrating both advantageous techniques, reducing treatment selection errors, and improving overall patient outcomes in CRC treatment.

Given these insights, this paper proposes an integrated approach to enhance treatment planning and patient outcomes in CRC care. By combining the powerful analytical capabilities of Machine Learning with the process insights of Process Mining, a more comprehensive and efficient medical decision support system can be created. More specifically, Process Mining can optimize treatment processes and improve patient care efficiency, while Machine Learning can provide accurate survival predictions and treatment response estimates, even considering subsequent treatment side effects and survival rate predictions. This combination not only helps physicians make more accurate clinical decisions but also provides patients with more personalized and efficient treatment options. As a result, from the synthesis of the literature and the proposed approach, the following research questions to solve the gaps are derived:

**Q1. How can Process Mining effectively support Machine Learning to optimize CRC treatment planning?**

- How can process mining be utilized to identify relevant features associated with CRC to assist in subsequent research?
- How can machine learning be utilized to predict treatment events while considering the influence of time series?

**Q2. What is the impact of this integrated approach on decision-making in CRC treatment?**

- What role does patient-specific data play in customizing treatment plans using this integrated approach?
- How can this approach aid physicians in making more informed decisions?

**Q3. What are the potential improvements in patient outcomes with the application of Process Mining and Machine Learning in CRC treatment?**

- How can this approach contribute to increasing the survival rates and quality of life of CRC patients?
- How can this technology be expanded to other fields in the future?

These research questions are designed to investigate how combining Process Mining and Machine Learning can improve CRC treatment planning. This exploration focuses on enhancing both the effectiveness and efficiency of the treatment strategies, which help to optimize decision-making processes, improve patient outcomes, and address the current gaps in treatment planning. This approach is expected to provide a more patient-centric and data-driven methodology that can be adapted to other cancer types in the future. The detailed research methods will be explained in subsequent sections.

# Chapter III

## Design Science Methodology

### 1 Engineering Cycle



FIGURE 3: Engineering cycle of the whole research

Through a systematic design science methodology[50], the structure and content of this research are clearly divided into stages. Viewing it as a complete Engineering cycle [FIGURE 3], it includes several steps: **Problem Investigation**, **Treatment Design**, **Treatment Validation**, **Treatment Implementation**, and finally, **Implementation Evaluation**. However, it is also a cyclical process, aimed at continuously optimizing and improving the solution.

In this study, **Chapters I and II** have already covered the Problem Investigation phase, identifying current research gaps and guiding toward potential solutions. The following **Chapters III to V** will detail how the treatment is designed and developed into a viable solution. Finally, in **Chapters VI to VIII**, specific validation methods will be designed to demonstrate the feasibility of this treatment and may show the directions for future improvement. Further, through a more systematic problem analysis, the following **Design Problem** can be summarized:

- Improve <the complex challenges of CRC treatment planning, including the medical, technological, and patient-specific factors,>
- By <establishing a system that integrates Process Mining and Machine Learning techniques for survival rate prediction>
- That satisfies <personalized strategies based on patient information,>
- In order to <provide doctors and patients with more informed and objective medical advice and quantitative bases.>

Once a solution emerges, it can further answer the research question of this research. At the same time, it provides a minimum viable model for subsequent Treatment Implementation to further expand the future applicability, including applications in different diseases or adding more functionalities. In the future, after the Implementation Evaluation in the real world, the model can be verified for its utility and potential improvements. This process not only demonstrates the effectiveness of the proposed solution in addressing the specific needs of colorectal cancer treatment planning but also lays the groundwork for its adaptation and enhancement for broader medical contexts.

## 2 Overview of the Method

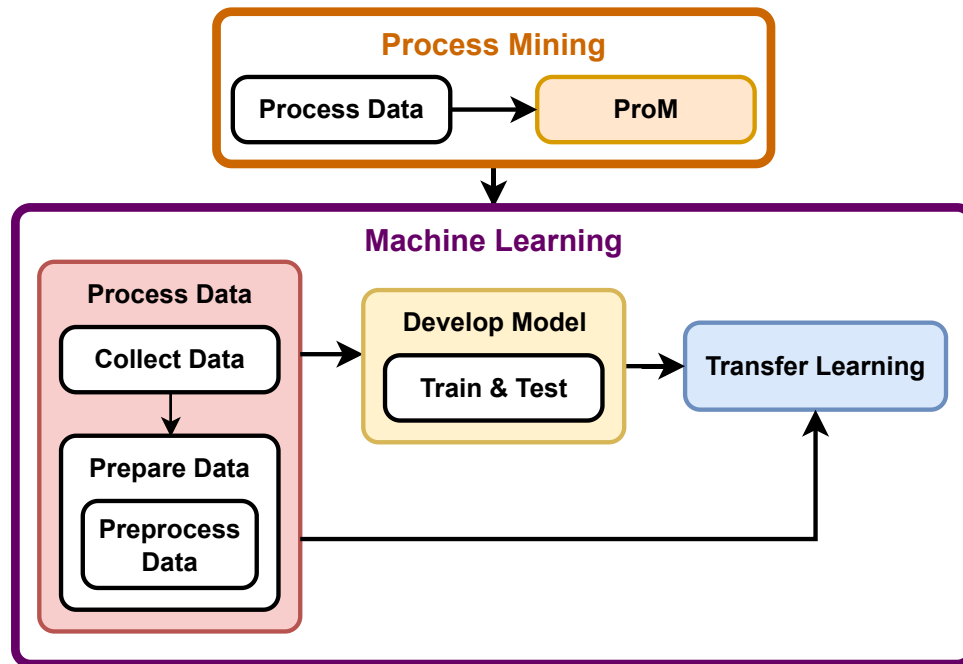


FIGURE 4: The overall process of the method

To answer the **Design Problem**, for the Methodology section, the process can be divided into two main flows [FIGURE 4]. After completing process mining, key feature information will be obtained, which can be effectively utilized in the Machine Learning model to produce valid outcomes. To be more specific, in the process mining phase, the focus is on the initial process data and the creation of diagrams using ProM software. This aims to extract relevant key event logs from the raw data and perform further feature extraction through qualitative analysis of the visualized process diagrams. Subsequently, the Machine Learning phase is divided into three steps: Process Data, Develop Model, and Transfer Learning. Data must first be converted into a computer-understandable format, then a model designed to predict the survival rate for each treatment is developed, and finally, transfer learning is employed to maximize data utilization and optimize model outcomes, which is also discovered to be feasible during the process mining. Through a series of methodological steps, it is possible to complete survival predictions for personalized treatment outcomes. The following sections will provide more detailed explanations of each part.



# Chapter IV

## Process Mining

### 1 Data Collection and Preprocessing

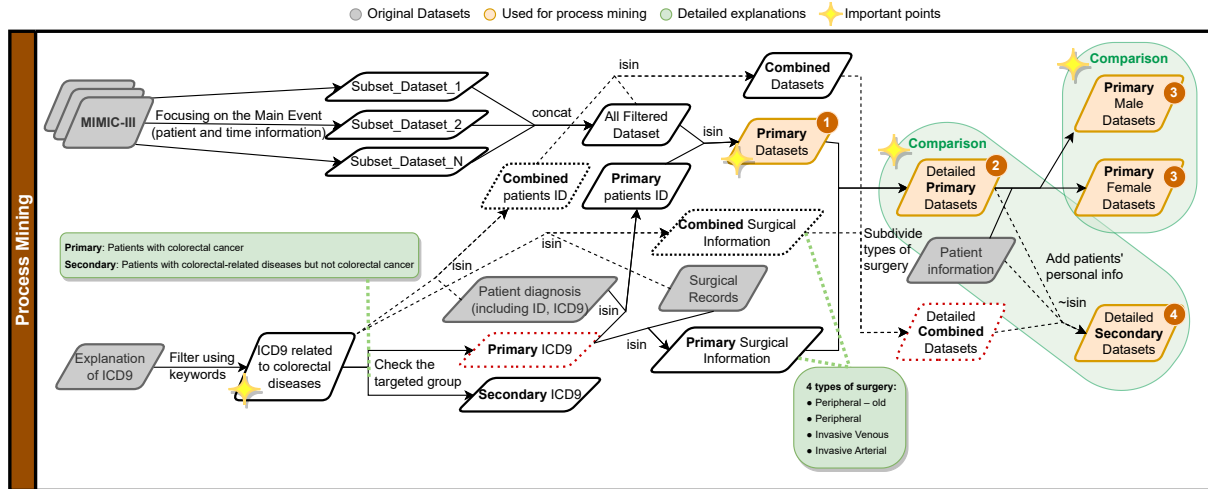


FIGURE 5: A detailed process mining flowchart

The *MIMIC-III* (Medical Information Mart for Intensive Care)[24] is a publicly available large medical database that contains information on over 40,000 patients from intensive care units at Israeli medical centers. The database includes patients' information, vital signs, medication, laboratory results, nursing records, imaging reports, encounter records, death times, etc. This provides detailed timestamps for Process Mining research, becoming essential pieces of the puzzle in constructing the entire process.

Following the arrows in [Figure 5], after a detailed review of the entire database, key extractions and integration of Event Logs containing patients, time, and event records are performed, resulting in a dataset named *All Filtered Dataset*. Additionally, it is necessary to utilize one of MIMIC's datasets named *Explanation of ICD9*, which encodes all diseases, to identify the targeted group. By searching for keywords such as "colon", "rectal", "colorectal", etc., diseases related to the colon are filtered out. Each ICD9 code is then individually confirmed for its relevance to the topic of colorectal cancer [Appendix B]. Given the smaller-than-expected total number of the targeted patients and to maximize the use of existing data, the remaining non-primary ones are retained and then both categorized into "Primary" and "Secondary" categories:

- **Primary:** Patients with Colorectal Cancer.
- **Secondary:** Patients with related colon diseases but not Colorectal Cancer.

With the ICD9 codes and *Patient diagnosis* information, the research subjects file named *Primary patients ID* can be used to filter the integrated dataset. Subsequently, the *Primary Datasets* are obtained for preliminary diagramming [Figure 5].

ProM, as a set of sophisticated tools and plugins, supports everything from basic process discovery to complex process enhancement and analysis, aiding in the visualization and improvement of actual business processes[54]. The software's main features include adopting diverse process mining algorithms, supporting multiple event log formats, and assisting users in comparing process executions under different times or conditions. Through the ProM tool's powerful diagramming capabilities, considering the Inductive Miner algorithm's strong noise tolerance and ability to handle complex processes, the process diagram results produced using Primary Datasets can be seen [Figure 6].

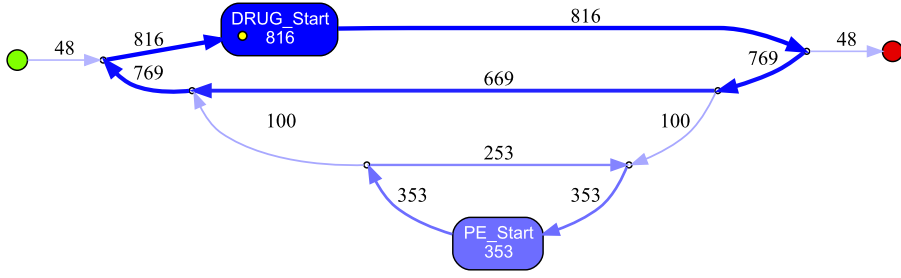


FIGURE 6: Initial PM Mapping Results (using Primary Datasets)

However, it is clear from the diagram that treatment options limited to medication (DRUG\_Start) and surgery (PE\_Start) are insufficient. By filtering through the *Primary Surgical Information* file and integrating it with the original dataset, the *Detailed Primary Datasets* are generated, which distinguish four categories of surgery [Figure 7]. This allows for a detailed process with 5 treatment options [Figure 8].

Surgery Type	Treatment Approach	Invasiveness or Treatment Intensity	Treatment Objectives
<b>Peripheral - old</b>	Non-invasive or minimally invasive surgery	Low	Reviewing old surgery sites, removing small tumors, or conducting adjunctive therapy
<b>Peripheral</b>	Minimally invasive surgery	Low to Medium	Establishing chemotherapy ports in peripheral body parts, or inspecting and treating distant metastases
<b>Invasive Venous</b>	Invasive surgery	Medium to High	Placing central venous catheters for chemotherapy, or excising large tumors near veins
<b>Invasive Arterial</b>	Major invasive surgery	High	Excising the main arteries supplying blood or performing extensive tumor resections

FIGURE 7: Explanations of the 4 Types of Surgery

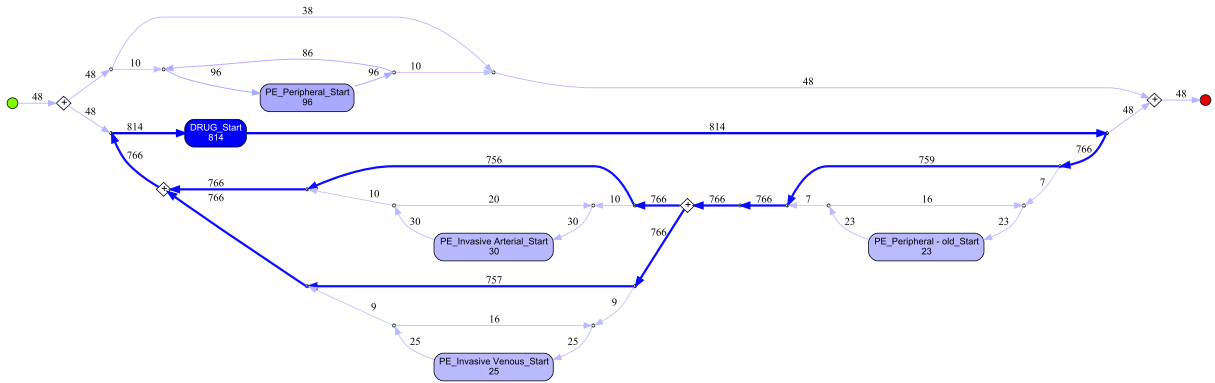


FIGURE 8: Second PM Mapping Results (using Detailed Primary Datasets)

Furthermore, through ProM’s process simplification tool, the main treatments can be identified as Peripheral-old, Peripheral, and Invasive Arterial [Figure 9]. Since Invasive Venous and Invasive Arterial are more similar, they will be considered in the same category. Along with medication (Drug\_Start), this leads to the identification of 4 main treatment options, which will also be the focus for further utilization of Machine Learning.

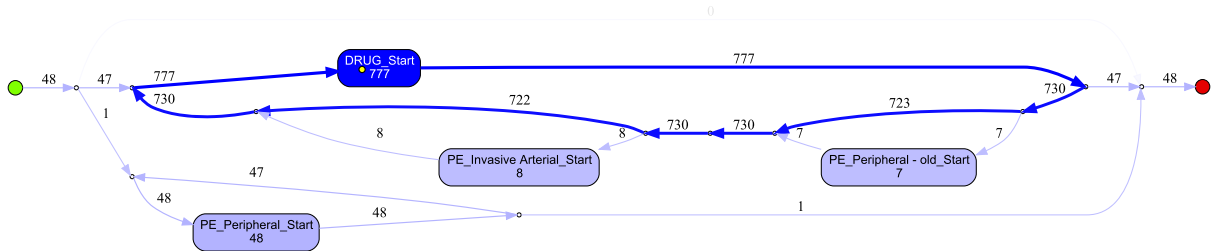


FIGURE 9: Simplified PM Mapping Results (using Detailed Primary Datasets)

## 2 Qualitative Analysis Using Process Mining

After identifying the primary treatment paths for target patients, it is possible to extract key patient features through more detailed categorization. When discussing colorectal cancer patients, age often serves as a basis for classification. The incidence rate of colorectal cancer increases with age. Based on age assessments, there exists a comprehensive system for dividing the impact of colorectal cancer across different age groups, especially for patients over 50 years old. Therefore, age is an obvious criterion for patient categorization. However, beyond age, by classifying patients by gender, creating **Primary Males/Female Datasets**, and conducting a qualitative analysis visually, it is observed that the treatment paths between males and females are not similar [Figure 10, 11]. Although there are fewer differences in cancer risk based on gender and the prevention and treatment applicability is higher (still limited to the four types of treatment methods), gender will also become one of the main features in building subsequent models for personalized treatment. This ensures that details within patient personal information are considered and applied, aiding in precise prediction.

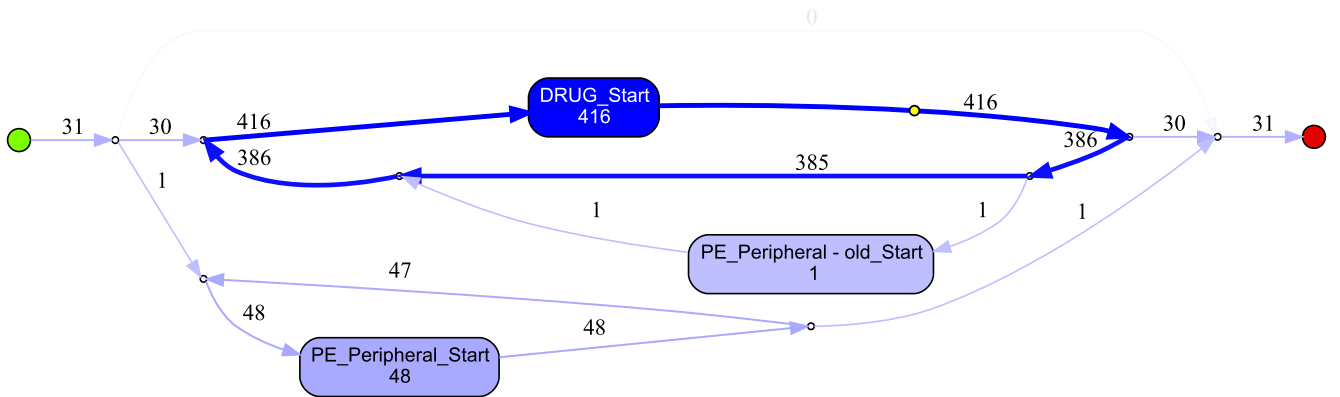


FIGURE 10: PM Mapping Results for Male (using Primary Male Datasets)

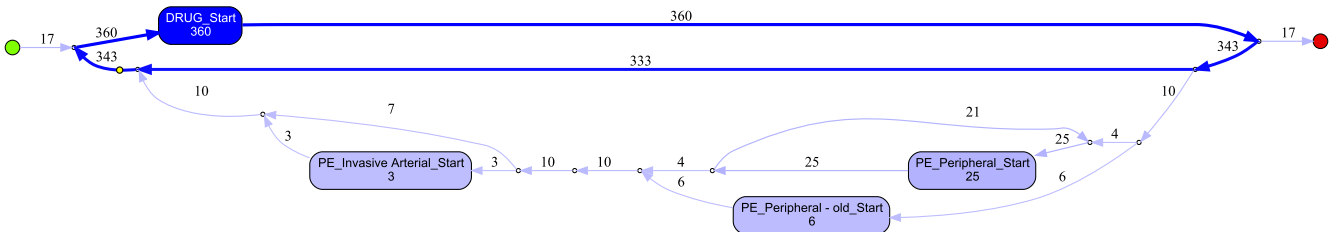


FIGURE 11: PM Mapping Results for Female (using Primary Female Datasets)

Moreover, due to the previously mentioned issue of insufficient data volume, by utilizing the Secondary category collected through the classified ICD9 codes and integrating it with the Primary to form **Detailed Combined Datasets**, the process data is conducted again in the same manner (following the dashed arrows in the diagram [Figure 5]), ultimately separating the files belonging to the Secondary category using a simple exclusion method. Independent files are then diagrammed using ProM [Figure 12, 13]. Upon comparison, it is discovered that the process output from the Primary can be found within the Secondary process. This suggests that the treatment process experienced by patients with colon-related diseases shares some similarities with that of colorectal cancer patients. This information will be further utilized in the machine learning segment.

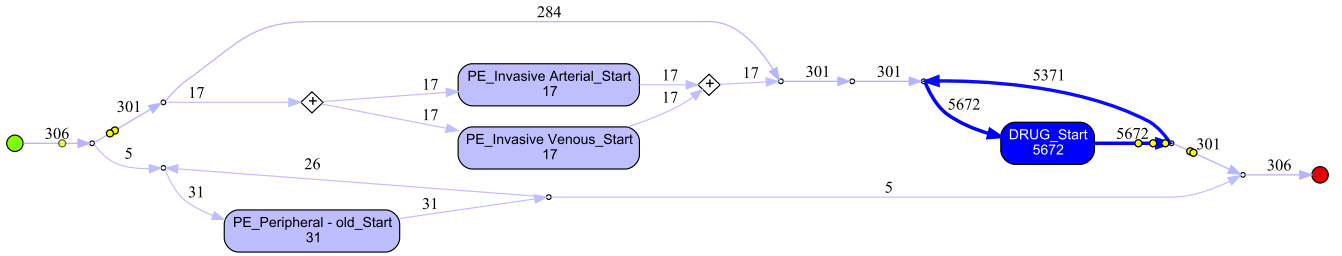


FIGURE 12: Primary PM Mapping Results (using Detailed Primary Datasets)

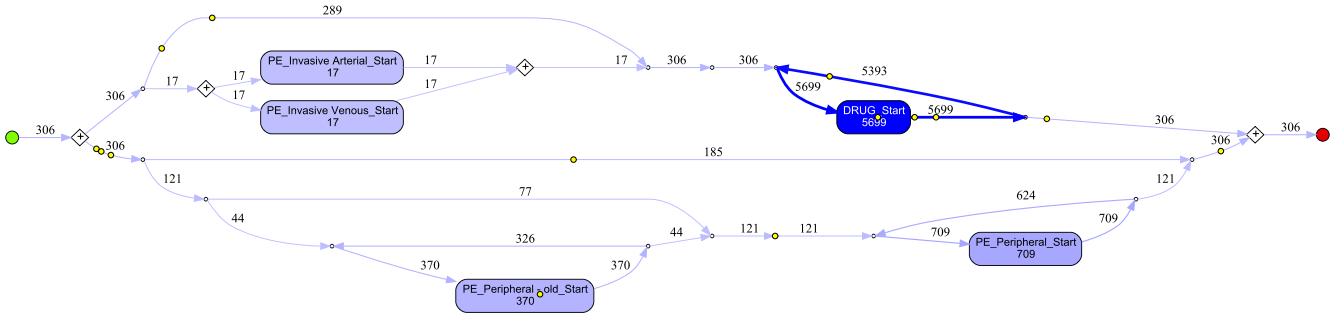


FIGURE 13: Secondary PM Mapping Results (using Detailed Secondary Datasets)

### 3 Summary

As a result, through process mining, the visualized diagrams of the treatment process help in identifying event logs related to the treatment among many events and in uncovering usable patient features that can be used in the model. Moreover, the analytical insights gained from process mining serve as a foundation for the facilitation of applying transfer learning techniques. This refers to the method of leveraging general knowledge gained from one task to another related task. This methodological finding introduces a new way to address the problem of data insufficiency, which can further enhance the model's precision and effectiveness.

# Chapter V

## Machine Learning

### 1 Data Feature Combined Extraction

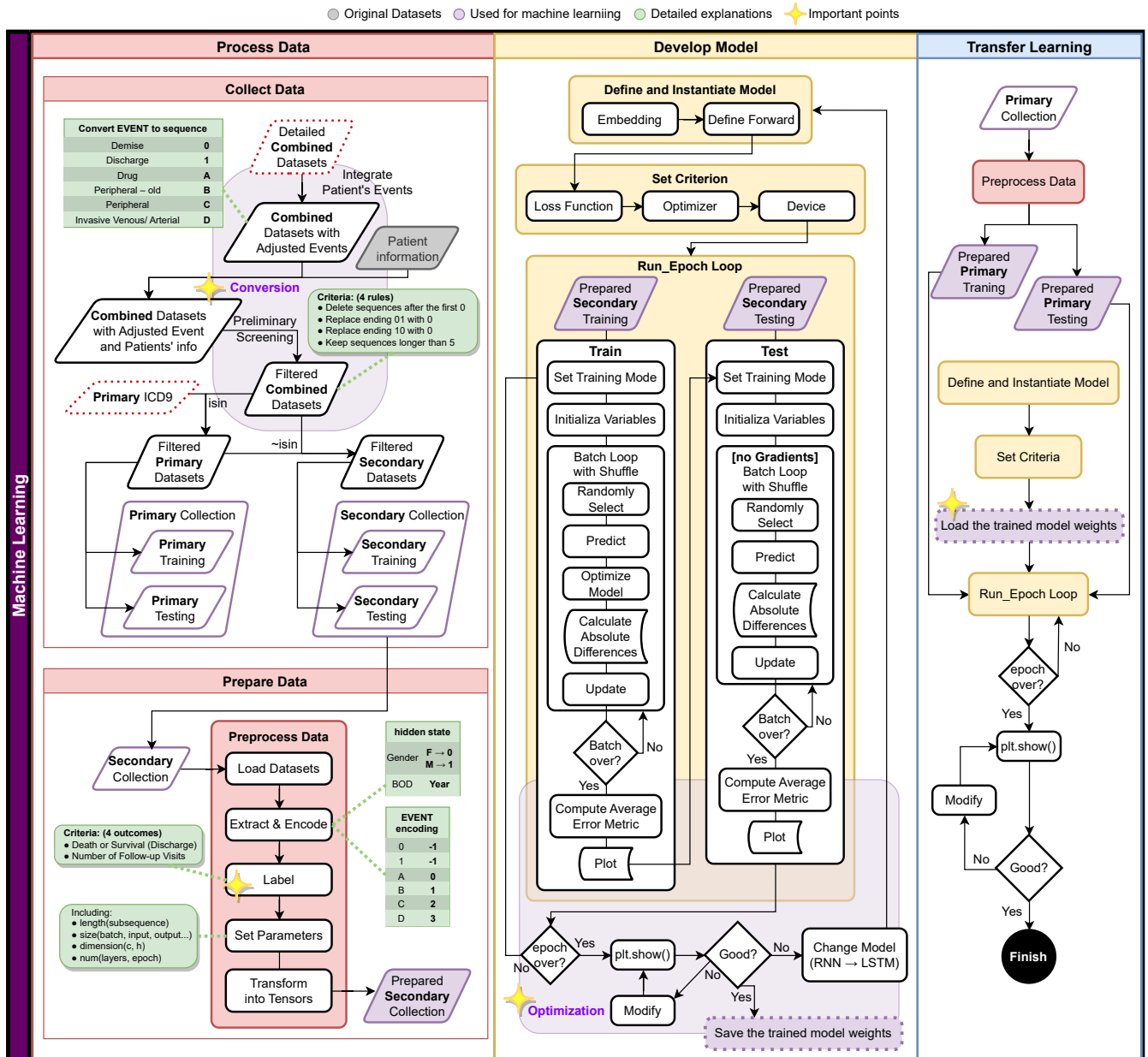



FIGURE 14: A detailed machine learning flowchart

Regarding the existing data, the qualitative analysis results from Process Mining reveal the 4 main treatment options and the key patient features, including age and gender. Moving to the machine learning part [Figure 14], by conducting preliminary event encoding on the *Detailed Combined Datasets* and combining it with *Patient Information*, a clearer patient event sequence file, named *Combined Datasets with Adjusted Event and Patients' info*, can be obtained. To elaborate, all events for each patient will be transformed into simplified codes and linked into a sequence arranged according to chronological order [Figure 15]. This step facilitates a manual screening before conversion into computer-understandable codes, helping to ensure the logical correctness of each sequence. Since patient survival is determined by discharge records, some sequences might end with records added posthumously, creating contradictions.

For example, if a patient dies in the hospital and is discharged afterward, the sequence’s end would be 01, necessitating preliminary adjustments to ensure the subsequent machine learning model does not encounter logical inconsistencies. Simultaneously, sequences shorter than five are excluded to ensure the model has sufficient information for understanding long-term dependencies and to prevent limited information from affecting or diminishing generalization capability. After preliminary screening and getting **Filtered Combined Datasets**, by using the previously filtered ICD9 code, the Primary and Secondary datasets can be split again for machine learning. Both are divided according to their data volume, thus 4 datasets are collected for subsequent model training and testing.

SUBJECT_ID	EVENT	TIME
1423	Peripheral	2180/8/3 18:30
1423	Discharge	2180/8/7 09:00
1423	Drug	2180/8/30 19:00
1423	Peripheral-old	2180/9/5 13:40
1423	Drug	2180/9/6 18:43
1423	Invasive	2180/9/7 13:30
1423	Demise	2180/9/10 05:10



SUBJECT_ID	EVENT
1423	C1ABAD0

Convert EVENT to sequence	
Demise	0
Discharge	1
Drug	A
Peripheral – old	B
Peripheral	C
Invasive	D

FIGURE 15: Preliminary Event Conversion Examples

Next, to make these data understandable to the computer, the extracted events and patient features need to be encoded, meaning they are translated into numerical form. Besides gender being simply encoded as 0 for females and 1 for males, it’s noteworthy that due to HIPAA regulations in MIMIC for patient privacy protection, some patient information, like birthdays, is encrypted and anonymized, hence only the year is extracted for the model to automatically find patterns.

Furthermore, the setting of labels is particularly crucial in the learning process of the model, affecting subsequent model performance and facilitating model interpretability. **The labeling method mainly uses two criteria: the patient’s final status and whether the patient returns for follow-up** [Figure 16]. The following provides detailed explanations for the selection of these indicators:

#### Patient’s Final Status:

- **Crucial for Predicting Survival Outcomes:** The ultimate goal of the model is to predict the survival outcomes of patients based on the various treatment plans they undergo. The final status of a patient (whether they survive or die) is a critical piece of information for this prediction.
- **Direct Impact on Model’s Accuracy:** The patient’s survival or death directly influences the model’s ability to accurately predict outcomes for similar future cases.
- **Informs Treatment Effectiveness:** The final status provides insight into the overall effectiveness of the treatment plan, guiding improvements in treatment strategies.

#### Patient Follow-up:

- **Evaluates Initial Treatment Results:** The frequency and necessity of patient follow-ups serve as an indirect measure of the initial treatment’s success. Frequent returns to the hospital indicate potential issues with the treatment’s effectiveness.



- **Adjustment in Labeling Process:** Follow-up data is used to adjust the labels in an arithmetic manner, reflecting the treatment process's significance and effectiveness. This approach is chosen over a geometric calculation to ensure that each treatment process is adequately valued.
- **Indicator of Treatment Efficacy:** A high number of follow-ups may suggest that the treatment was not fully effective, prompting the model to adjust its predictions and potentially identify less effective treatment methods.
- **Incorporation into Model Learning:** The information from follow-ups is integrated into the model's learning process, allowing it to consider the efficacy of initial treatments and their impact on patient outcomes.

To be more specific, the reason for not using a geometric calculation is to ensure the significance of each treatment process. Imagine if a patient dies after less than two treatments, the model might strongly deduce a severe fault in a certain treatment method during training. However, many poor outcomes result from accumulated mistakes in the early process or the patient's own health condition. Therefore, this experiment assumes that any treatment measure has a certain effect on the patient but will infer and evaluate the survival results of intermediate treatments based on the final outcome arithmetically. As seen in the example [Figure 16], the final results are also verified in code form to ensure labels are correctly represented under different circumstances. As a result, by focusing on these two criteria, the model aims to provide a more accurate and nuanced understanding of treatment outcomes, enhancing its predictive capability and aiding in the improvement of treatment strategies.

		Patient's Final Status	
		1 (Discharged /Survived)	0 (Demised)
Number of hospital admissions throughout the entire event (occurrences of 1)	Only once	Ex. AAAA <b>1</b> A → 1 A → 1 A → 1 A → 1 1	Ex. AAAAA <b>0</b> A → 0.8 A → 0.6 A → 0.4 A → 0.2 A → 0 0
	more than once	Ex. AAA <b>1</b> AAAA <b>1</b> A → 0.8 A → 0.8 A → 0.8 1 A → 0.85 A → 0.9 A → 0.95 A → 1 1	Ex. AAA <b>1</b> AAAA <b>0</b> A → 0.6 A → 0.6 A → 0.6 1 A → 0.4 A → 0.2 A → 0 0

Ex. EVENT: AAB**B**CC**1**DD**1**AA**1**CC**0**  
 → 0.6, 0.6, 0.6, 0.6, 0.6, 0.6, 0.6, 0.6, 0.6, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1, 0  
 ([ 'A', 'A', 'B', 'B', 'C', 'C', 'D', 'D', 'D', 'A', 'A', 'A', 'C', 'C', 'D' ],  
 [0.6, 0.6, 0.6, 0.6, 0.6, 0.6, 0.6, 0.6, 0.6, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1, 0.0])

FIGURE 16: The 4 Labeling Strategies

Additionally, model parameters need to be set, including the sequence length during training, various sizes (batch, input, output), dimensions, and quantities (layers, epochs, etc.). Finally, to run the data on PyTorch and operate it on a GPU, it needs to be converted into tensors, thus completing the data preparation. Later, the model will undergo hyperparameter adjustment based on the results.

## 2 Machine Learning Model Development and Training

The model development process includes initially defining and instantiating the model, and setting up the criterion to consider an appropriate loss function and optimizer [Figure 14]. The overall model architecture, in addition to the basic processes and weight optimization during training and testing, also incorporates a special shuffle method. Since the training sequence length (subsequence) and batch have already been set initially, how to prevent the model from "forgetting" by avoiding sequential data retrieval needs consideration. To allow the model to randomly access information from different time stages within each sequence, an infinite loop is established that runs continuously until there is insufficient data to form a complete batch. In each iteration, the code randomly selects a set of patient information through the *random.sample* function, which simultaneously extracts training results and hidden states from the previous step [Figure 17]. More specifically, this method disrupts the original order of the data. For each selected sample, the code also extracts corresponding features and labels according to chronological order from the dataset, such as gender, birth of date, and previously occurred events, assembling them into new batches. This method of randomly selecting samples effectively prevents gradient explosion or disappearance during model training due to specific time periods, forcing it to learn more general and essential data characteristics, thereby improving predictive performance.

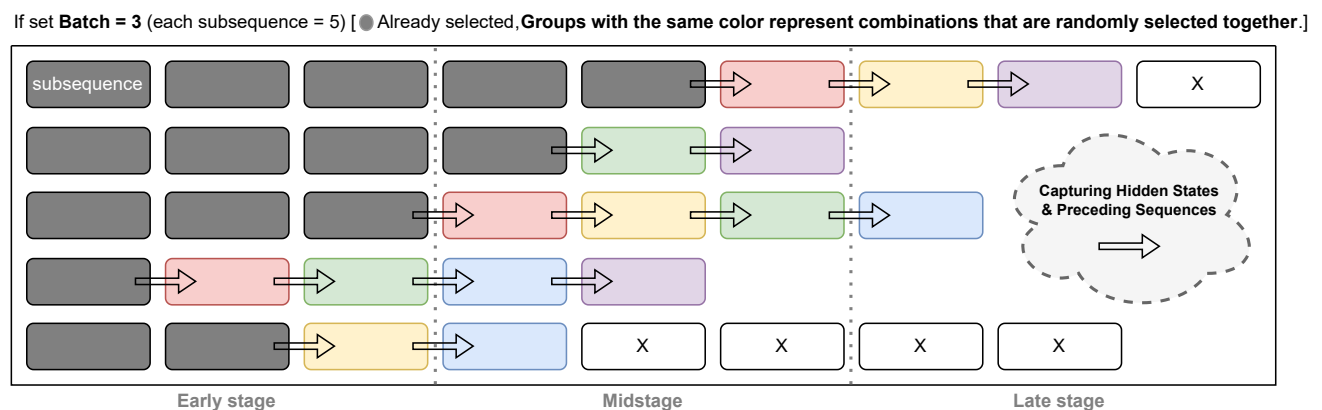


FIGURE 17: An illustration of shuffling

After running all epochs, the results are optimized using methods like hyperparameter adjustment and model changes. Initially, an **RNN(Recurrent Neural Network)** was chosen as the primary model for the current research direction [Figure 14]. In the field of predictive healthcare research, it has been proven that RNNs can be employed to analyze large-scale historical electronic health record (EHR) data, predicting future medical conditions and medication requirements with high accuracy and generalizability across various institutions[13]. As a neural network adept at handling sequential data, RNNs can capture information in time series using hidden states (h) [Figure 18], which are updated over time. Additionally, through the use of the tanh function, RNNs offer good gradient properties and help compress information to deal with complex data patterns and relationships, making it the primary choice for preliminary model testing.



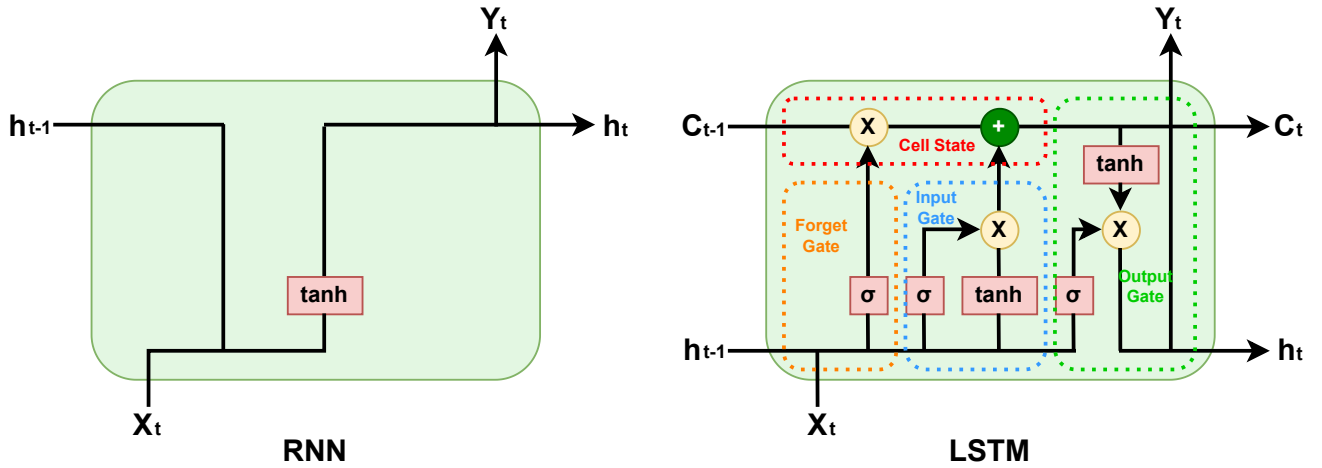


FIGURE 18: Model Comparison between RNN and LSTM

However, subsequent optimization revealed the limitations of RNN models, namely, their difficulty in capturing long-term dependencies. Numerous studies have highlighted similar issues[10]. Besides, despite attempts to adjust hyperparameters, the results did not significantly change. It's estimated that as the sequence length increases, gradients are still prone to explosion or disappearance, which is a problem still faced despite the special shuffle method. Therefore, during optimization, attempts were made to change the model to a special type of RNN model, **LSTM(Long Short-Term Memory)**, and readjust the model internals [Figure 18]. LSTM, by introducing a three-gate mechanism: **Forget Gate**, **Input Gate**, **Output Gate**, working alongside the **Cell State**, overcomes the challenges of long sequences. The Forget Gate enables LSTM to effectively add or delete crucial information to the Cell State, ensuring vital information remains and better learning long-term dependencies. When the Cell state combines important information with the current input, it also goes through the tanh function for information compression, eventually becoming the output ( $Y=h$ ) and potentially saving crucial information ( $C$ ) from the Output Gate.

Overall, compared to RNNs, while LSTM structures are more complex and require more computational resources, they offer improved performance for handling complex long-sequence tasks and dependencies. Thus, after continuous modifications and model adjustments, a generally trained model weight can ultimately be obtained [Figure 14].

### 3 Transfer Learning

Transfer learning is a machine learning technique that allows the knowledge learned from one task to be applied to another related task [Figure 19]. In past applications within medical research, it is utilized in scenarios with limited data availability by leveraging pre-trained models on extensive datasets and fine-tuning them on smaller, specific datasets, considerably improving medical image classification tasks despite the initial scarcity of data[7]. In other words, pre-trained models have usually learned a rich and general set of features and patterns from their original tasks. By transferring this knowledge, the model can perform better on new tasks, especially in situations where the data volume is small. Moreover, based on the results of Process Mining, since the process information of the Primary is contained within the Secondary, it supports further utilization of Transfer Learning for secondary optimization of the pre-trained model [Figure 12, 13]. The overall process is similar to the previous step [Figure 14], but it is important to note that the training and testing sets of Primary should be used. By saving the trained model weights and loading them in a new environment, the weights can be optimized using the target dataset, ultimately achieving more accurate prediction results.

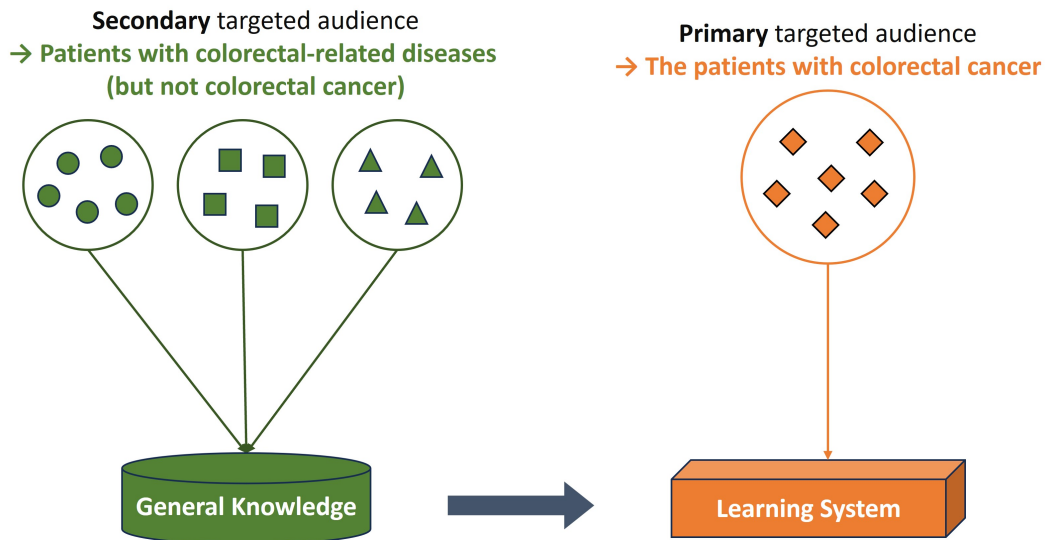


FIGURE 19: The Explanation of Transfer Learning

## 4 Summary

To summarize, by processing the dataset integrated from process mining and further filtering the event sequence complied with specific rules, and then applying a specialized labeling method, the data can be input into a constructed model for preliminary training and testing. Additionally, to further optimize the results, transfer learning is employed to increase the data volume and enhance the model's predictive accuracy. Ultimately, based on different treatment options and patient information, survival outcome predictions with confidence intervals are provided to the patients, which can help in making an informed decision.

# Chapter VI

## Experimental Validation

### 1 Explanation of the evaluation formula

After the model has been trained and tested, to evaluate the outcomes of the model's operation, an **Error Metric (EM)** is set as the criterion for judgment. The formula is defined as follows:

$$\mathbf{EM} = |\mathbf{P} - \mathbf{L}| \quad (\text{VI.1})$$

(where P is the prediction of survival rate, L is the label of each treatment.)

Each subsequence will accumulate the prediction difference for each event using the formula below to calculate a total value: (*total\_abs\_diff += torch.sum(abs\_diff).item()*)

$$\mathbf{Total\ EM\ of\ Each\ Subsequence} = \sum_{i=1}^n |\mathbf{P}_i - \mathbf{L}_i| \quad (\text{VI.2})$$

(where P is the prediction of survival rate, L is the label of each treatment, n is the length of the subsequence.)

After an epoch is completed, the total number of updates to the "Total EM of Each Subsequence" is computed as the denominator, representing the total event length of all patients (*filtered\_num\_samples += filtered\_y\_tensor.numel()*). It is important to note that due to the special shuffle method mentioned above, when a batch is insufficient to form a complete combination, the epoch will end, hence the total sequence length will vary each time. Additionally, to ensure each subsequence is fully filled, "-1" was used for padding; however, when calculating the total number of events used, these "-1" values need to be excluded. Therefore, the total number of events summed is filtered. Thus, at the end of each epoch, an average Error Metric serves as the final judgment criterion, with the formula as follows:

$$\mathbf{Average\ EM\ of\ Each\ Epoch} = \frac{\sum_{i=1}^m \left( \sum_{j=1}^n |\mathbf{P}_{ij} - \mathbf{L}_{ij}| \right)}{\mathbf{num\_events}} \quad (\text{VI.3})$$

(where P is the prediction of survival rate, L is the label of each treatment, n is the length of the subsequence, m is the total number of the batch, num\_events is the total used num of all patients' events.)

Therefore, the result calculated represents that the lower the Avg\_EM (Average EM of Each Epoch), the better its prediction outcome and the higher the accuracy; vice versa. It also represents the deviation from the actual correct results, which introduces a concept of the confidence interval. Subsequent visualizations produced by the model will provide further detailed explanations of the results.

## 2 Model Results

Through continuous hyperparameter adjustments, the RNN model, after being trained and tested through the set epochs, finally reached an Avg\_EM of **0.2** for training and **0.28** for testing [Figure 20]. The graph intuitively shows that as epochs increase over time and weights are optimized within the model, the Avg\_EM shows a downward trend, eventually converging to the aforementioned results. This indicates that the learning situation is good and parameters within the model are continuously optimized. Moreover, comparing the trends of training and testing, both exhibit similar trends so there is no overfitting or underfitting phenomena.

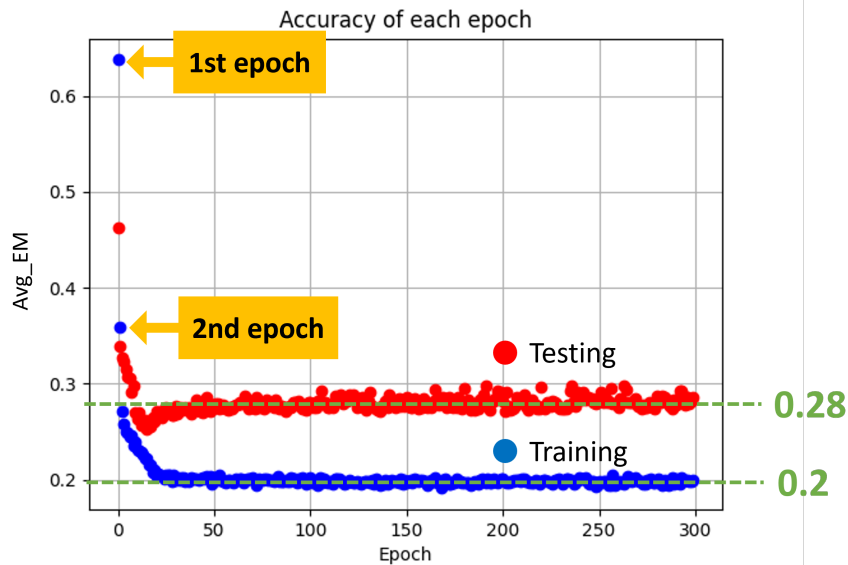


FIGURE 20: RNN evaluation results

However, at the current stage, an error value of 0.2 as a result is still not ideal. Therefore, after adjusting the RNN model's hyperparameters to the point where it converged without further reducing the Avg\_EM, the model was changed to LSTM. Ultimately, the Avg\_EM for training remained at **0.2**, but testing slightly decreased to **0.26** [Figure 21].

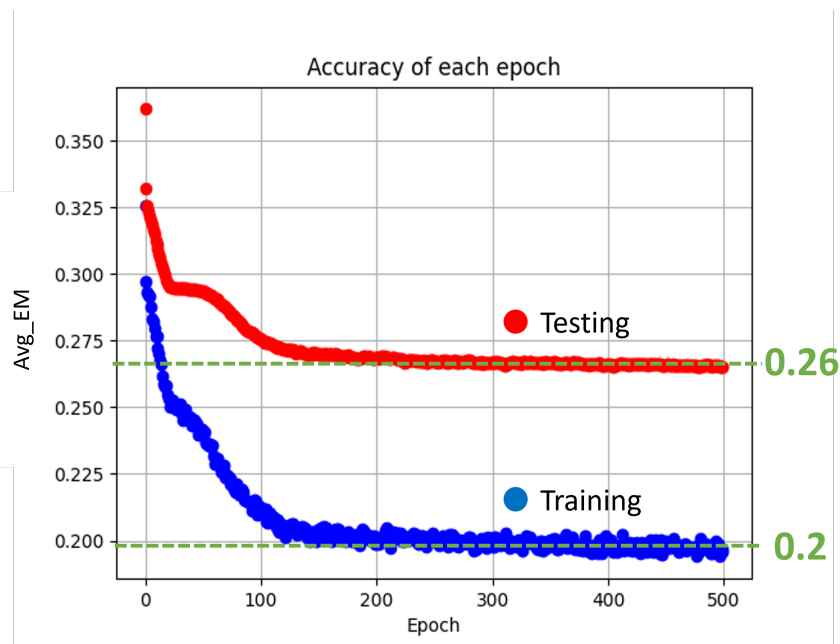


FIGURE 21: LSTM evaluation results

Nevertheless, although there was a slight improvement in results, after adjustment, they converged to roughly the same outcome. Hence, reconsidering the condition of the data itself, it is speculated that the possible reason might be the excessive diversity in the dataset. This speculation arises because the dataset adopted earlier was from the Secondary patient group (*Prepared Secondary Collection* in Figure 14), so the current model results may have only learned more general knowledge. After employing the *Prepared Primary Collection* for Transfer Learning, the Avg\_EM for training remained at **0.2**, but testing significantly dropped to **0.13** [Figure 22]. The trend of the red dots improved by **0.13** and performed better than the blue, which is a better result than expected.

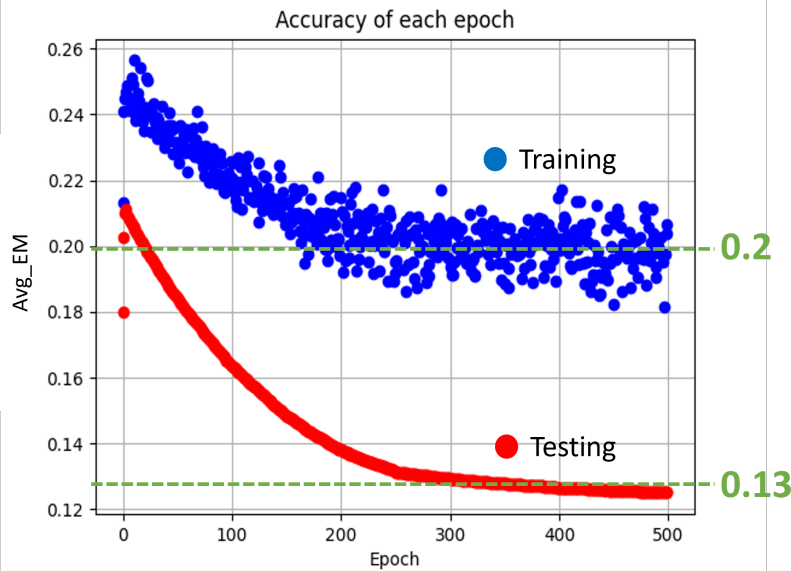


FIGURE 22: Evaluation results after Transfer Learning

To ensure this result was not due to the coincidental use of an appropriate dataset, the *Prepared Primary Collection* was re-split for training and testing. The Avg\_EM for training dropped to **0.178**, while the Avg\_EM for testing was **0.17** [Figure 23]. Compared to the previous results, they are more reasonable and the two trends are closer. Overall, after repeated optimization, model conversion, and implementing Transfer Learning, the final Error Metric averaged around **0.15** (taking the average of two results).

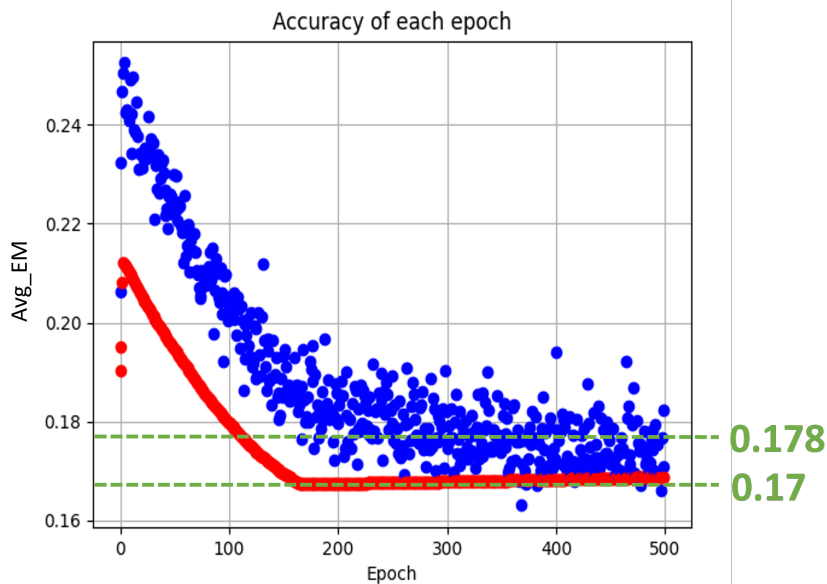


FIGURE 23: Evaluation results after re-splitting the dataset and adopting Transfer Learning

### 3 Summary

By repeatedly splitting the primary datasets for training and testing, and comparing both extreme and reasonable outcomes, it's observed that the accuracy of the model still improves with the number of epochs. This confirms the model's feasibility. Moreover, by averaging both results, an approximate error metric of 0.15 is achieved. This means that future predictions will have a confidence interval, allowing predicted values to vary within a plus or minus 0.15 range. However, the results produced by the model enable patients to choose the most suitable treatment method by directly comparing the highest survival rates.

# Chapter VII

## Discussion

In this study, we explored the application of process mining and machine learning technologies in the treatment planning of colorectal cancer (CRC), aiming to predict patient survival rates through personalized treatment plans and provide quantifiable decision-making indicators. By leveraging medical data from the MIMIC database and the powerful analytical capabilities of ProM software, we successfully identified key event logs related to CRC treatment and trained a machine learning model based on these feature pieces of information. To be more specific, the final results of the model, for instance, predicting a survival rate of 70% (represented as 0.7), provide a confidence interval of  $\pm 0.14$  (Avg\_EM), indicating that the actual survival rate is likely between 56% and 84%. This research offers a quantitative measure of treatment effectiveness uncertainty to doctors and patients, aiding them in making more informed decisions. Additionally, as the error values for different treatment options are consistent, direct input and comparison of the predicted outcomes of the four treatment options can help to determine the best treatment method for the patient as well. Below, we systematically answer the research questions in this article and propose specific expansion strategies:

### A1. Integrating Process Mining and Machine Learning for CRC Treatment

The application of process mining technology in this study highlights its role in process discovery within CRC treatment planning. By analyzing open-source databases, we extracted event logs related to CRC and discovered gender differences in treatment through visualization, aiding and validating the feasibility of subsequent Transfer Learning with comparisons between primary and secondary patient groups. This provided the machine learning model with rich feature information and extended research directions. The methods section also discussed the ability and feasibility of machine learning in predicting treatment events when dealing with time series data, bringing new perspectives and insights into CRC treatment prediction.

### A2. Impacting CRC Decision-Making through Integrated Approaches

The integrated approach adopted in this study significantly enhanced the accuracy and personalization of CRC treatment decisions. By offering predicted survival rates and confidence intervals to doctors and patients, this research supports treatment choices based on more comprehensive information. Moreover, the model also considered individualized data, such as age and gender, further promoting personalized customization of treatment plans, and enabling doctors to devise more precise and effective treatment strategies for patients. On the other hand, it also assists patients and their families, with lack expert knowledge, in making treatment decisions with quantitative benchmarks tailored to their conditions, avoiding any subjective or norm-based erroneous decisions.

### A3. Enhancing CRC Patient Outcomes with Advanced Analytics

This study demonstrates the potential of applying process mining and machine learning in CRC treatment. Not only can it provide accurate treatment prediction results, but it also has the potential to further uncover key features, thereby significantly improving patient treatment outcomes. As technology advances, this approach could be extended to a wider range of medical fields, using data from other cancers or diseases, to provide personalized treatment plans for various illnesses.

Moreover, it's important to note that despite the achievements of this study, there are some limitations. Data insufficiency is one of the main challenges, as the performance of machine learning models largely depends on a vast amount of high-quality training data. To overcome this challenge and achieve better optimization in the future, research could consider the following directions:

- **Data Integration:** Combining clinical data from multiple hospitals to increase data volume and diversity, improving the model's generalizability and accuracy.
- **Feature Enrichment:** Introducing more patient information as the model's hidden states, such as family medical history, personal medical history, dietary habits, BMI, and lifestyle habits, to provide a more comprehensive view of the patient's health status.
- **Technological Updates:** With medical technology advancement, new treatment methods and technologies continuously emerge, necessitating ongoing updates and adjustments to the model to reflect the latest treatment options and technological levels. Regularly optimizing the model to include the latest medical knowledge and treatment methods is key to ensuring the model's long-term effectiveness.

Through these improvements, based on the methods of this article, it hopes to further enhance the accuracy of personalized treatment plan prediction technology in the future, providing stronger support and assistance for achieving truly personalized medicine.



# Chapter VIII

## Conclusion and Future Work

This study conducted an in-depth analysis of the treatment planning for colorectal cancer through the integration of Process Mining and Machine Learning technologies, proposing a more objective methodology. By conducting a detailed analysis of medical data from MIMIC-III and employing advanced ProM analysis tools, key event logs related to CRC treatment were successfully identified. Through qualitative analysis of visualized process diagrams, the identified feature information was utilized to train a Machine Learning model, ultimately achieving effective predictions of personalized treatment plans and providing a path for future research in feature selection.

Moreover, practically, this study offers tangible benefits to healthcare professionals involved in the planning and administration of colorectal cancer treatments. The utilization of this integrated approach provides a more objective and data-driven methodology for treatment planning, which specifically aids medical practitioners by offering insights into efficient treatment strategies and enhancing the personalized care for colorectal cancer patients. Hospitals and clinical institutions can also leverage the findings of this research to improve their treatment planning systems, making use of the predictive capabilities of the developed Machine Learning model. This not only improves the quality of patient care but also serves as a sample for incorporating advanced data analysis in the decision-making processes, paving the way for more informed and effective treatment methodologies in clinical practice. To further expand the applicability of this technology, the following aspects could be considered:

- **Interdisciplinary Collaboration:** Collaborating with experts from various fields can help address the issue of insufficient data, while also enhancing the accuracy and generalizability of the model. For instance, integrating with bioinformatics can provide deeper insights into diseases; advancing together with data science can optimize or develop more efficient algorithms for handling big data; and incorporating statistical knowledge can aid in assessing the accuracy and credibility of model predictions. Such interdisciplinary collaborations can better handle complex medical data and consider more factors to make more accurate predictions.
- **Data Sharing Platform:** Establishing a secure data-sharing platform with compliance and permissions can encourage different hospitals and research institutions to share medical data. This not only solves the problem of data insufficiency but also enhances data diversity, allowing the model to more accurately reflect the treatment responses of different populations.
- **Ethical Considerations:** Relevant policies and principles need to be established to ensure the protection of patient privacy and rights when using Machine Learning models for medical decision-making, while also preventing data leakage or exposure. Innovations driven by data should also ensure that no individual's rights are infringed upon during the sharing of resources.

In conclusion, by adopting these methods, we can further enhance the accuracy, usability, and reliability of personalized treatment plan prediction technology, providing strong support for achieving truly personalized medicine. This will not only improve patient treatment outcomes and quality of life but also bring more efficient and cost-effective treatment solutions to the healthcare system.

# Appendix A

Reference	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Total Score
[5]	0	5	2	3	4	3	2	19
[41]	3	0	0	0	3	2	1	9
[49]	3	3	2	2	4	3	2	19
[40]	1	1	0	4	5	5	4	20
[42]	0	1	1	1	2	3	1	9
[23]	0	1	0	2	5	4	2	14
[18]	0	0	2	2	2	2	1	9
[26]	0	0	0	3	2	1	5	11
[45]	0	3	0	1	4	4	3	15
[1]	0	0	0	3	2	1	5	11
[17]	0	3	0	2	3	3	3	14
[6]	0	3	0	4	5	4	4	20
[14]	0	5	0	4	4	3	4	20
[28]	0	0	2	3	1	0	3	9
[35]	0	5	2	5	5	4	4	25
[2]	0	5	0	2	5	3	3	18
[20]	0	5	0	3	5	3	4	20
[8]	0	0	0	3	3	1	3	10
[32]	0	3	0	4	5	2	3	17
[9]	3	1	0	3	4	3	5	19
[25]	0	5	0	4	5	4	3	21
[3]	2	3	0	3	3	2	5	18
[29]	0	5	3	3	4	3	4	22
[48]	3	1	0	3	2	2	2	13
[47]	3	1	0	3	4	3	2	16
[27]	0	5	3	2	4	3	3	20
[21]	0	0	0	3	2	1	2	8
[33]	2	1	0	3	2	1	2	11
[55]	0	3	0	3	4	2	3	15
[16]	0	0	0	3	3	0	2	8
[34]	0	5	0	4	4	2	3	18
[51]	0	0	1	2	4	3	2	12
[11]	0	3	0	3	3	3	2	14
[4]	1	5	0	4	3	3	3	19
[43]	0	3	0	3	3	2	3	14
[15]	0	0	0	3	2	2	1	8
[46]	0	1	0	3	4	2	1	11
[36]	0	3	0	3	3	3	2	14
[22]	3	1	0	4	4	2	4	18
[30]	0	3	0	4	4	4	4	19
[44]	1	1	3	3	3	1	3	15
[52]	0	3	0	3	3	3	4	16

TABLE A: Detailed Quality Assessment According to 7 Scoring Criteria

# Appendix B

Colorectal Cancer	ICD9_CODE	English Name
O	1492	Mal neo transverse colon
O	1493	Mal neo descend colon
O	1494	Mal neo sigmoid colon
O	1497	Malig neo ascend colon
O	1499	Malignant neo colon NEC
O	1500	Malignant neo colon NOS
O	2171	Ca in situ colon
O	2623	Mal crcnoid ascend colon
O	2624	Mal crcnoid transv colon
O	2625	Mal carcinoid desc colon
O	2626	Mal carcinoid sig colon
X	2651	Ben carcinoid asc colon
X	2652	Ben crcinoid trans colon
X	2653	Ben carcinoid desc colon
X	2654	Ben carcinoid sig colon
X	4945	Dvrtelo colon w/o hmrhg
X	5188	Pseudopolyposis colon
X	5216	Dvrtcli colon w/o hmrhg
X	5217	Dvrtelo colon w hmrhg
X	5218	Dvrtcli colon w hmrhg
X	5229	Megacolon NEC
X	5235	Anal & rectal abscess
X	5254	Anal & rectal polyp
X	5259	Anal or rectal pain
X	9860	Hx of colonic malignancy
X	9861	Hx-rectal & anal malign
X	10875	Screen malig neop-colon
X	11724	FB in intestine & colon
X	12151	Ascending colon inj-clos
X	12152	Transverse colon inj-cl
X	12153	Descending colon inj-cl
X	12154	Sigmoid colon inj-closed
X	12159	Ascending colon inj-open
X	12160	Transverse colon inj-opn
X	12161	Descending colon inj-opn
X	12162	Sigmoid colon inj-open
X	12973	Family hx colonic polyps
X	13383	Prsnl hst colonic polyps

TABLE B: Detailed ICD-9 Classification Results

# Reference

- [1] M.M. Abdelsamea, U. Zidan, Z. Senousy, M.M. Gaber, E. Rakha, and M. Ilyas. A survey on artificial intelligence in histopathology image analysis. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(6), 2022. doi: [10.1002/widm.1474](https://doi.org/10.1002/widm.1474).
- [2] S. Ahmad, T. Ullah, I. Ahmad, A. Al-Sharabi, K. Ullah, R.A. Khan, S. Rasheed, I. Ullah, M.N. Uddin, and M.S. Ali. A novel hybrid deep learning model for metastatic cancer detection. *Computational Intelligence and Neuroscience*, 2022, 2022. doi: [10.1155/2022/8141530](https://doi.org/10.1155/2022/8141530).
- [3] Z. Al-Taie, D. Liu, J.B. Mitchem, C. Papageorgiou, J.T. Kaifi, W.C. Warren, and C.-R. Shyu. Explainable artificial intelligence in high-throughput drug repositioning for subgroup stratifications with interventionable potential. *Journal of Biomedical Informatics*, 118, 2021. doi:[10.1016/j.jbi.2021.103792](https://doi.org/10.1016/j.jbi.2021.103792).
- [4] N.S. Alghamdi. Evaluation of classification models for predicting mortality rate using thyroid cancer data. *Journal of Computer Science*, 15(1):131–142, 2019. doi: [10.3844/jcssp.2019.131.142](https://doi.org/10.3844/jcssp.2019.131.142).
- [5] F.A. Altuhaifa, K.T. Win, and G. Su. Predicting lung cancer survival based on clinical data using machine learning: A review. *Computers in Biology and Medicine*, 165, 2023. doi:[10.1016/j.combiomed.2023.107338](https://doi.org/10.1016/j.combiomed.2023.107338).
- [6] M. Alwohaibi, M. Alzaqebah, N.M. Alotaibi, A.M. Alzahrani, and M. Zouch. A hybrid multi-stage learning technique based on brain storming optimization algorithm for breast cancer recurrence prediction. *Journal of King Saud University - Computer and Information Sciences*, 34(8):5192–5203, 2022. doi:[10.1016/j.jksuci.2021.05.004](https://doi.org/10.1016/j.jksuci.2021.05.004).
- [7] Laith Alzubaidi, Muthana Al-Amidie, Ahmed Al-Asadi, Amjad J. Humaidi, Omran Al-Shamma, Mohammed A. Fadhel, Jinglan Zhang, J. Santamaría, and Ye Duan. Novel transfer learning approach for medical imaging with limited labeled data. *Cancers*, 13(7):1590, 2021. Number: 7 Publisher: Multidisciplinary Digital Publishing Institute. URL: <https://www.mdpi.com/2072-6694/13/7/1590>, doi:[10.3390/cancers13071590](https://doi.org/10.3390/cancers13071590).
- [8] A. Aminifar, M. Shokri, F. Rabbi, V.K.I. Pun, and Y. Lamo. Extremely randomized trees with privacy preservation for distributed structured health data. *IEEE Access*, 10:6010–6027, 2022. doi:[10.1109/ACCESS.2022.3141709](https://doi.org/10.1109/ACCESS.2022.3141709).
- [9] P. Asghari. A diagnostic prediction model for colorectal cancer in elderlies via internet of medical things. *International Journal of Information Technology (Singapore)*, 13(4):1423–1429, 2021. doi:[10.1007/s41870-021-00663-5](https://doi.org/10.1007/s41870-021-00663-5).
- [10] Merijn Beeksma, Suzan Verberne, Antal van den Bosch, Enny Das, Iris Hendrickx, and Stef Groenewoud. Predicting life expectancy with a long short-term memory recurrent neural network using electronic medical records. *BMC Medical Informatics and Decision Making*, 19(1):36, 2019. doi:[10.1186/s12911-019-0775-2](https://doi.org/10.1186/s12911-019-0775-2).
- [11] C. Blake and R. Kehm. Comparing breast cancer treatments using automatically detected surrogate and clinically relevant outcomes entities from text. *Journal of Biomedical Informatics: X*, 1, 2019. doi:[10.1016/j.yjbinx.2019.100005](https://doi.org/10.1016/j.yjbinx.2019.100005).

- [12] Faiza Allah Bukhsh, Zaharah Allah Bukhsh, and Maya Daneva. A systematic literature review on requirement prioritization techniques and their empirical evaluation. *Computer Standards & Interfaces*, 69:103389, 2020. URL: <https://www.sciencedirect.com/science/article/pii/S0920548919300789>, doi:10.1016/j.csi.2019.103389.
- [13] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. Doctor AI: Predicting clinical events via recurrent neural networks. In *Proceedings of the 1st Machine Learning for Healthcare Conference*, pages 301–318. PMLR, 2016. ISSN: 1938-7228. URL: <https://proceedings.mlr.press/v56/Choi16.html>.
- [14] A.Z. Dag, Z. Akcam, E. Kibis, S. Simsek, and D. Delen. A probabilistic data analytics methodology based on bayesian belief network for predicting and understanding breast cancer survival. *Knowledge-Based Systems*, 242, 2022. doi:10.1016/j.knosys.2022.108407.
- [15] E.M.F. El Houbay. A survey on applying machine learning techniques for management of diseases. *Journal of Applied Biomedicine*, 16(3):165–174, 2018. doi:10.1016/j.jab.2018.01.002.
- [16] N. Fatima, L. Liu, S. Hong, and H. Ahmed. Prediction of breast cancer, comparative review of machine learning techniques, and their analysis. *IEEE Access*, 8:150360–150376, 2020. doi:10.1109/ACCESS.2020.3016715.
- [17] M. Field, D.I. Thwaites, M. Carolan, G.P. Delaney, J. Lehmann, J. Sykes, S. Vinod, and L. Holloway. Infrastructure platform for privacy-preserving distributed machine learning development of computer-assisted theragnostics in cancer. *Journal of Biomedical Informatics*, 134, 2022. doi:10.1016/j.jbi.2022.104181.
- [18] L.H.A. Fryan and M.B. Alazzam. Survival analysis of oncological patients using machine learning method. *Healthcare (Switzerland)*, 11(1), 2023. doi:10.3390/healthcare11010080.
- [19] Anne-Wil Harzing and Satu Alakangas. Google scholar, scopus and the web of science: a longitudinal and cross-disciplinary comparison. *Scientometrics*, 106(2):787–804, 2016. doi:10.1007/s11192-015-1798-9.
- [20] T. He, J. Li, P. Wang, and Z. Zhang. Artificial intelligence predictive system of individual survival rate for lung adenocarcinoma. *Computational and Structural Biotechnology Journal*, 20:2352–2359, 2022. doi:10.1016/j.csbj.2022.05.005.
- [21] A.M. Hemeida, S. Alkhalaf, A. Mady, E.A. Mahmoud, M.E. Hussein, and A.M. Baha Eldin. Implementation of nature-inspired optimization algorithms in some data mining tasks. *Ain Shams Engineering Journal*, 11(2):309–318, 2020. doi:10.1016/j.asej.2019.10.003.
- [22] M. Hoogendoorn, P. Szolovits, L.M.G. Moons, and M.E. Numans. Utilizing uncoded consultation notes from electronic medical records for predictive modeling of colorectal cancer. *Artificial Intelligence in Medicine*, 69:53–61, 2016. doi:10.1016/j.artmed.2016.03.003.
- [23] E.K. Jadoon, F.G. Khan, S. Shah, A. Khan, and M. Elaffendi. Deep learning-based multi-modal ensemble classification approach for human breast cancer prognosis. *IEEE Access*, 11:85760–85769, 2023. doi:10.1109/ACCESS.2023.3304242.

- [24] Alistair Johnson, Tom Pollard, and Roger Mark. MIMIC-III Clinical Database, 2015. URL: <https://physionet.org/content/mimiciii/1.4/>, doi:10.13026/C2XW26.
- [25] K. Karami, M. Akbari, M.-T. Moradi, B. Soleymani, and H. Fallahi. Survival prognostic factors in patients with acute myeloid leukemia using machine learning techniques. *PLoS ONE*, 16(7), 2021. doi:10.1371/journal.pone.0254976.
- [26] R. Kaul, C. Ossai, A.R.M. Forkan, P.P. Jayaraman, J. Zelcer, S. Vaughan, and N. Wickramasinghe. The role of AI for developing digital twins in healthcare: The case of cancer care. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 13(1), 2023. doi:10.1002/widm.1480.
- [27] I. Kaur, M.N. Doja, and T. Ahmad. Time-range based sequential mining for survival prediction in prostate cancer. *Journal of Biomedical Informatics*, 110, 2020. doi:10.1016/j.jbi.2020.103550.
- [28] I. Kaur, M.N. Doja, and T. Ahmad. Data mining and machine learning in cancer survival research: An overview and future recommendations. *Journal of Biomedical Informatics*, 128, 2022. doi:10.1016/j.jbi.2022.104026.
- [29] I. Kaur, M.N. Doja, T. Ahmad, M. Ahmad, A. Hussain, A. Nadeem, and A.A. Abd El-Latif. An integrated approach for cancer survival prediction using data mining techniques. *Computational Intelligence and Neuroscience*, 2021, 2021. doi:10.1155/2021/6342226.
- [30] W. Kim, K.S. Kim, and R.W. Park. Nomogram of naive bayesian model for recurrence prediction of breast cancer. *Healthcare Informatics Research*, 22(2):89–94, 2016. doi:10.4258/hir.2016.22.2.89.
- [31] Barbara Ann Kitchenham, David Budgen, and Pearl Brereton. *Evidence-Based Software Engineering and Systematic Reviews*. CRC Press, 2015. Google-Books-ID: bGfdCgAAQBAJ.
- [32] D. Liu, J. Shao, H. Liu, and W. Cheng. Design on early warning system for renal cancer recurrence based on CNN-based internet of things. *IEEE Access*, 10:34835–34845, 2022. doi:10.1109/ACCESS.2021.3114227.
- [33] M. Loey, M.W. Jasim, H.M. EL-Bakry, M.H.N. Taha, and N.E.M. Khalifa. Breast and colon cancer classification from gene expression profiles using data mining techniques. *Symmetry*, 12(3), 2020. doi:10.3390/sym12030408.
- [34] G. López-García, J.M. Jerez, L. Franco, and F.J. Veredas. Transfer learning with convolutional neural networks for cancer survival prediction using gene-expression data. *PLoS ONE*, 15(3), 2020. doi:10.1371/journal.pone.0230536.
- [35] Y. Ma, H. Zhu, Z. Yang, and D. Wang. Optimizing the prognostic model of cervical cancer based on artificial intelligence algorithm and data mining technology. *Wireless Communications and Mobile Computing*, 2022, 2022. doi:10.1155/2022/5908686.
- [36] H. Min, H. Mobahi, K. Irvin, S. Avramovic, and J. Wojtusiak. Predicting activities of daily living for cancer patients using an ontology-guided machine learning methodology. *Journal of Biomedical Semantics*, 8(1), 2017. doi:10.1186/s13326-017-0149-6.

- [37] Philippe Mongeon and Adèle Paul-Hus. The journal coverage of web of science and scopus: a comparative analysis. *Scientometrics*, 106(1):213–228, 2016. doi:[10.1007/s11192-015-1765-5](https://doi.org/10.1007/s11192-015-1765-5).
- [38] Inés Mármol, Cristina Sánchez-de Diego, Alberto Pradilla Dieste, Elena Cerrada, and María Jesús Rodríguez Yoldi. Colorectal carcinoma: A general overview and future perspectives in colorectal cancer. *International Journal of Molecular Sciences*, 18(1):197, 2017. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute. URL: <https://www.mdpi.com/1422-0067/18/1/197>, doi:[10.3390/ijms18010197](https://doi.org/10.3390/ijms18010197).
- [39] Jessica B O’Connell, Melinda A Maggard, Edward H Livingston, and Clifford K Yo. Colorectal cancer in the young. *The American Journal of Surgery*, 187(3):343–348, 2004. URL: <https://www.sciencedirect.com/science/article/pii/S0002961003005981>, doi:[10.1016/j.amjsurg.2003.12.020](https://doi.org/10.1016/j.amjsurg.2003.12.020).
- [40] A. Panigrahi, A. Pati, B. Sahu, M.N. Das, D.S.K. Nayak, G. Sahoo, and S. Kant. En-MinWhale: An ensemble approach based on MRMR and whale optimization for cancer diagnosis. *IEEE Access*, 11:113526–113542, 2023. doi:[10.1109/ACCESS.2023.3318261](https://doi.org/10.1109/ACCESS.2023.3318261).
- [41] N. Qarmiche, K. El Kinany, N. Otmani, K. El Rhazi, and N.E.H. Chaoui. Cluster analysis of dietary patterns associated with colorectal cancer derived from a moroccan case-control study. *BMJ Health and Care Informatics*, 30(1), 2023. doi:[10.1136/bmjhci-2022-100710](https://doi.org/10.1136/bmjhci-2022-100710).
- [42] P. Ramakrishna and P. Rajarajeswari. Evolutionary optimization algorithm for classification of microarray datasets with mayfly and whale survival. *International journal of online and biomedical engineering*, 19(13):17–37, 2023. doi:[10.3991/ijoe.v19i13.40145](https://doi.org/10.3991/ijoe.v19i13.40145).
- [43] A.N. Richter and T.M. Khoshgoftaar. A review of statistical and machine learning methods for modeling cancer risk using structured clinical data. *Artificial Intelligence in Medicine*, 90:1–14, 2018. doi:[10.1016/j.artmed.2018.06.002](https://doi.org/10.1016/j.artmed.2018.06.002).
- [44] A. Sharma, J. Hostetter, J. Morrison, K. Wang, and E. Siegel. Focused decision support: a data mining tool to query the prostate, lung, colorectal, and ovarian cancer screening trial dataset and guide screening management for the individual patient. *Journal of Digital Imaging*, 29(2):160–164, 2016. doi:[10.1007/s10278-015-9826-0](https://doi.org/10.1007/s10278-015-9826-0).
- [45] A. Sheidaei, A.R. Foroushani, K. Gohari, and H. Zeraati. A novel dynamic bayesian network approach for data mining and survival data analysis. *BMC Medical Informatics and Decision Making*, 22(1), 2022. doi:[10.1186/s12911-022-02000-7](https://doi.org/10.1186/s12911-022-02000-7).
- [46] N. Shukla, M. Hagenbuchner, K.T. Win, and J. Yang. Breast cancer data analysis for survivability studies and prediction. *Computer Methods and Programs in Biomedicine*, 155:199–208, 2018. doi:[10.1016/j.cmpb.2017.12.011](https://doi.org/10.1016/j.cmpb.2017.12.011).
- [47] D. Sui, K. Zhang, W. Liu, J. Chen, X. Ma, and Z. Tian. CST: A multitask learning framework for colorectal cancer region mining based on transformer. *BioMed Research International*, 2021, 2021. doi:[10.1155/2021/6207964](https://doi.org/10.1155/2021/6207964).
- [48] P. Thareja and R.S. Chhillar. Comparative analysis of data mining algorithms for cancer gene expression data. *International Journal of Advanced Computer Science and Applications*, 12(10):322–328, 2021. doi:[10.14569/IJACSA.2021.0121035](https://doi.org/10.14569/IJACSA.2021.0121035).



- [49] C. Tudor and R.A. Sova. Mining google trends data for nowcasting and forecasting colorectal cancer (CRC) prevalence. *PeerJ Computer Science*, 9, 2023. doi:10.7717/PEERJ-CS.1518.
- [50] R. J. Wieringa. *Design Science Methodology for Information Systems and Software Engineering*. Springer, 2014. URL: [https://books.google.com/books/about/Design\\_Science\\_Methodology\\_for\\_Informati.html?hl=zh-TW&id=xLKLQBQAQBAJ](https://books.google.com/books/about/Design_Science_Methodology_for_Informati.html?hl=zh-TW&id=xLKLQBQAQBAJ).
- [51] X. Wu, H. Akbarzadeh Khorshidi, U. Aickelin, Z. Edib, and M. Peate. Imputation techniques on missing values in breast cancer treatment and fertility data. *Health Information Science and Systems*, 7(1), 2019. doi:10.1007/s13755-019-0082-4.
- [52] R. Xu and Q. Wang. Large-scale automatic extraction of side effects associated with targeted anticancer drugs from full-text oncological articles. *Journal of Biomedical Informatics*, 55:64–72, 2015. doi:10.1016/j.jbi.2015.03.009.
- [53] Lai Xue, Ashley Williamson, Sara Gaines, Ciro Andolfi, Terrah Paul-Olson, Anu Neerukonda, Emily Steinhagen, Radhika Smith, Lisa M. Cannon, Blasé Polite, Konstantin Umanskiy, and Neil Hyman. An Update on Colorectal Cancer. *Current Problems in Surgery*, 55(3):76–116, March 2018. URL: <https://www.sciencedirect.com/science/article/pii/S0011384018300248>, doi:10.1067/j.cpsurg.2018.02.003.
- [54] Fitri Almira Yasmin, Rob Bemthuis, Moustafa Elhagaly, Fons Wijnhoven, and Faiza Allah Bukhsh. A process mining starting guideline for process analysts and process owners: A practical process analytics guide using ProM. *DSI technical report series*, 2020. URL: <https://research.utwente.nl/en/publications/a-process-mining-starting-guideline-for-process-analysts-and-proc>.
- [55] Z.M. Zain, M. Alshenaifi, A. Aljaloud, T. Albednah, R. Alghanim, A. Alqifari, and A. Alqahtani. Predicting breast cancer recurrence using principal component analysis as feature extraction: An unbiased comparative analysis. *International Journal of Advances in Intelligent Informatics*, 6(3):313–327, 2020. doi:10.26555/ijain.v6i3.462.