UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

MEDICAL CLASSIFICATION THROUGH MACHINE LEARNING, DOMAIN-SPECIFIC LANGUAGE MODELS AND GENERATIVE AI TECHNIQUES

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

BY

REZA GHEIBI
Norman, Oklahoma
2026

MEDICAL CLASSIFICATION THROUGH MACHINE LEARNING,
DOMAIN-SPECIFIC LANGUAGE MODELS AND GENERATIVE AI TECHNIQUES

A DISSERTATION APPROVED FOR THE
SCHOOL OF COMPUTER SCIENCE

BY THE COMMITTEE CONSISTING OF

Dr. Dean Frederick Hougen, chair

Dr. Ronald Barnes

Dr. Sanjana Mudduluru

Dr. Egawati Panjei

Acknowledgements

I want to begin by expressing my gratitude to my advisor, Dr. Dean Hougen, for his guidance, support, and encouragement throughout my Ph.D. journey. His expertise, insights, and patience have been invaluable in shaping my research and enabling me to achieve my goals.

The author sincerely thanks the members of the research committee, Dr. Ronald Barnes, Dr. Sanjana Mudduluru, and Dr. Egawati Panjei. I am honored to have benefited from their mentorship, which has significantly shaped my development as a researcher.

My thanks also go to my colleagues and friends for their patience and understanding throughout this intensive process. I greatly appreciate your support during periods of absence and look forward to reconnecting with each of them in the near future.

Finally, I must acknowledge the unwavering support of my family. Their constant encouragement and belief in my abilities were instrumental in overcoming the challenges of this journey; I share the successful completion of this degree with all of you.

Abstract

Advancing reliable machine learning and natural language processing methods for classification under data constrained conditions remains a central challenge in computer science. This dissertation addresses core problems related to data scarcity, class imbalance, heterogeneous tabular features, and the integration of structured and unstructured representations. To investigate these challenges, cervical cancer classification is used as an application domain where limited, noisy, and imbalanced datasets make model development particularly difficult.

The work first evaluates traditional machine learning models including logistic regression, decision trees, random forests, gradient boosting, and multilayer perceptrons under varying data quality conditions. Multiple data augmentation strategies are examined, such as SMOTE, generative adversarial networks, diffusion-based models, large language models, domain-specific small language models, and synthetic tabular data generation, demonstrating that appropriate augmentation can substantially improve robustness and predictive performance.

The second component introduces a method for transforming structured medical records into natural language representations, allowing the use of modern language models for classification. Domain specific and general purpose transformers (e.g., PubMedBERT, BioBERT, GPT-2, LLaMA, Mistral) are fine tuned and evaluated in supervised, zero-shot, and few-

shot settings. The results show that text-based modeling can match or exceed traditional machine learning performance, particularly in low-resource scenarios, while capturing richer contextual information.

Finally, the dissertation extends this language-based methodology to a large-scale clinical taxonomy task. Through unsupervised analysis and systematic refinement, ambiguous specialty labels are consolidated into coherent Super-Specialties, and a fine-tuned PubMedBERT classifier achieves strong multi-label performance. This demonstrates the scalability of the proposed methods for organizing complex clinical data.

In general, this dissertation demonstrates that combining structured data modeling, generative data augmentation, and language-based representations provides a flexible and effective framework for cervical cancer classification under data-constrained conditions. The findings highlight the potential of these approaches to support more accessible, scalable, and robust clinical decision-support systems, with broader implications for medical classification tasks beyond cervical cancer.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter introduces the broader computational challenges associated with developing reliable AI systems for medical data analysis and establishes the critical need for predictive machine learning models, which are often hindered by limitations in conventional screening methods, including accessibility and subjective interpretation. To address these challenges, the chapter motivates the application of artificial intelligence (AI) and machine learning (ML) to improve diagnostics. However, it identifies significant barriers to effective ML deployment, namely data scarcity, class imbalance, and privacy constraints. Consequently, the chapter proposes data augmentation, synthetic data generation via generative AI, and innovative tabular-to-text transformations using language models as core strategies to overcome these data limitations. Finally, it outlines the research objectives centered on developing and evaluating these data-driven frameworks and presents the organizational structure of the dissertation.

1.1 Artificial Intelligence and Machine Learning in Medical Diagnosis

Artificial intelligence and machine learning have emerged as powerful tools to analyze complex medical data in a wide range of clinical applications. In medical research, ML-based approaches have been applied to various data, including images, histopathological slides, genomic data, and structured patient records, demonstrating promising improvements in diagnostic accuracy and efficiency (Litjens et al., 2017; Hou et al., 2022; Wang et al., 2024).

Traditional ML models, such as logistic regression, decision trees, random forests, and gradient boosting, offer interpretable and computationally efficient solutions for structured clinical data analysis. More recently, deep learning architectures have enabled automated feature extraction from high-dimensional inputs, further expanding the scope of AI-assisted diagnostics (LeCun et al., 2015). However, the performance of these models is highly dependent on the availability of large, high-quality labeled datasets, which are often difficult to obtain in medical domains (Rajkomar et al., 2018).

Despite these advances, the deployment of AI systems in clinical practice remains challenging due to concerns related to data quality, generalizability, interpretability, and regulatory compliance (El Arab et al., 2025). These challenges are particularly pronounced in cervical cancer datasets, which are often small, imbalanced, and heterogeneous (Litjens et al., 2017; Hou et al., 2022).

1.2 Data Scarcity, Imbalance, and Privacy Challenges

One of the most significant barriers to effective classification of medical data is the limited availability of high-quality annotated datasets. Medical data collection is constrained by

ethical considerations, privacy regulations, and the cost of expert annotation (El Arab et al., 2025). As a result, datasets often contain relatively few samples, with a substantial class imbalance between healthy and diseased cases.

Class imbalance poses a major challenge for supervised learning algorithms, which may become biased toward majority classes and exhibit poor sensitivity for minority or high-risk cases (Salmi et al., 2024). In clinical contexts, such failures can have serious consequences, as misclassifications of positive cases undermines early detection efforts (Johnson and Khoshgoftaar, 2019).

Privacy regulations governing personal health information further restrict data sharing and reuse. Although these regulations are essential for protecting patient confidentiality, they limit opportunities for large-scale data aggregation and cross-institutional collaboration (Matta and Bolli, 2025). Consequently, researchers must develop methods that can operate effectively under data-constrained conditions while preserving patient privacy (Price and Cohen, 2019).

1.3 Motivation for Data Augmentation and Generative AI

Data augmentation and synthetic data generation have gained increasing attention as strategies to mitigate data scarcity and imbalance in medical AI (Fayaz et al., 2024). Data augmentation techniques aim to increase the diversity of the data set by generating modified versions of existing samples, thereby reducing overfitting and improving model generalization (Shorten and Khoshgoftaar, 2019). In structured clinical datasets, enhancement methods, such as resampling, noise injection, and feature perturbation, can help balance class distributions and improve robustness (Gurcan and Soylu, 2024; Vanitha and Kasthuri, 2025).

Synthetic data generation offers a more comprehensive solution by creating entirely new samples that reflect the statistical properties of real data. Generative adversarial networks (GANs), diffusion models, and other generative frameworks have demonstrated the ability to produce realistic medical data while preserving privacy (Frid-Adar et al., 2018; Ho et al., 2020). These approaches enable researchers to expand datasets without exposing sensitive patient information and facilitate more extensive experimentation under controlled conditions.

This dissertation systematically evaluates multiple augmentation and generative techniques, examining their impact on classification performance and robustness in cervical cancer datasets as a case study.

1.4 Language Models and Tabular-to-Text Transformation

Although traditional ML models are well suited for structured data, recent advances in natural language processing have opened new opportunities for medical data analysis. Many clinical records contain rich contextual information that is difficult to capture using purely numerical representations. Transforming structured tabular data into natural language descriptions enables the application of transformer-based language models that excel at capturing contextual and semantic relationships.

Domain-specific language models, such as PubMedBERT and BioBERT, have demonstrated strong performance on biomedical NLP tasks by pretraining in large-scale scientific corpora. General-purpose models, including GPT-based architectures, LLaMA, and Mistral also offer capabilities such as zero-shot and few-shot learning, which are particularly valuable in low-data settings (Brown et al., 2020; Lee et al., 2020; Gu et al., 2021; Harvard Medical

School, 2023).

The tabular-to-text paradigm explored in this dissertation leverages these strengths by reformulating patient records as textual narratives, enabling both classification and patient-centered communication. This approach provides a flexible framework for integrating structured data with contextual reasoning while maintaining computational efficiency through the use of smaller, fine-tuned models (Kale and Rastogi, 2020; Hegselmann et al., 2023).

1.5 Case Study: Cervical Cancer Dataset Classification

According to the World Health Organization (WHO) (World Health Organization, 2024), cervical cancer develops in a cervix (the entrance to the uterus from the vagina). Cancer can often spread throughout the body. Cervical cancer is simply the abnormal growth of body cells that forms in organ tissues that connect the upper part of the uterus and the vagina (birth canal), called the *cervix*. Cervical cancer is the most common cause of female genital cancer and female cancer deaths in developing countries such as Nigeria. It is the second most common type of cancer in Nigerian women and the leading gynecological malignancy with high mortality among the afflicted (Oluwole et al., 2017). If not detected and treated early, cervical cancer is nearly always fatal. Almost all cervical cancer cases (99%) are linked to infection with high-risk human papilloma-viruses (HPV), an extremely common virus transmitted through sexual contact. Although most HPV infections resolve spontaneously and do not cause symptoms, persistent infection can cause cervical cancer in women. In addition, cervical cancer is the fourth most common cancer among women. In 2018, an estimated 570,000 women were diagnosed with cervical cancer worldwide, and approximately 311,000 women died from the disease, giving a mortality rate of approximately 55%. Effective primary prevention approaches (HPV vaccination) and secondary prevention approaches (screening and treatment of precancerous lesions) will prevent most cases of

cervical cancer. When diagnosed, cervical cancer is one of the most successfully treatable forms of cancer if it is detected early and effectively managed. Early detection greatly improves the chances of successful treatment of pre-cancers and cancer. Estimation of risk and awareness of any signs and symptoms of cervical cancer can also help avoid delays in diagnosis. Especially, when cervical cancer is diagnosed earlier, death can be avoided. Screening tests offer the best chance to have cervical cancer detected early, when treatment can be most successful. Screening can also actually prevent most cervical cancers by finding abnormal cervical cell changes (pre-cancers) so that they can be treated before they have a chance to turn into a cervical cancer. Cervical cancer screening is a valuable early detection tool that can identify cancer in an early stage when treatment has been proven to be more effective and less expensive (Arbyn et al., 2020; Cao et al., 2024).

Hence, the importance of the development of machine learning and artificial intelligence techniques cannot be overemphasized, as such systems can help support early risk assessment for cervical cancer by identifying individuals who may be at elevated risk and should therefore be referred for further clinical examination. This helps save time, mental, financial, and emotional stress that can be caused by cervical cancer. However, the accuracy of these diagnostic models can sometimes be compromised due to discrepancies in the datasets used to train them. These discrepancies can lead to instances of incorrect diagnoses, which undermine the effectiveness of these tools (Aljrees, 2024; Muraru et al., 2024). This dissertation highlights the common discrepancies found in such datasets, explores their impact on the performance of machine learning models, and discusses how these inaccuracies can ultimately affect the reliability of cervical cancer diagnoses.

Limited health awareness and delayed health-seeking behaviors remain significant challenges in many populations. Cervical cancer is no exception: numerous women remain unaware of their personal risk levels due to limited access to screening services and inade-

quate dissemination of relevant health information. In rural communities in Southwestern Nigeria, for example, women and girls are often isolated from critical health knowledge due to a combination of poverty, geographic distance, and limited healthcare resources. When access to timely screening and preventive services is limited, preventable diseases can continue to progress undetected, regardless of the awareness or intentions of the individuals (Denny et al., 2017; Shin et al., 2021).

To address these communication and accessibility challenges, this dissertation explores how machine learning, natural language processing, and generative AI can be used to transform basic user-provided information into an understandable context-appropriate health guide. AI models can process input such as age, reproductive history, and other non-clinical demographic factors to generate general information related risk that users can comprehend. Thus, improving awareness and supporting informed decision making.

Consequently, when information on cervical cancer (risk estimation and awareness) is communicated to these women and girls, it will go a long way toward preserving their lives and community because they will have a better understanding of cervical cancer and its associated risks. Having a good understanding of the signs and symptoms of cervical cancer can also help avoid delayed diagnosis. Early detection of the disease allows for faster action and greatly improves the chances of successful treatment.

With the increase in health data in terms of the capacity and availability, the role of machine learning in diagnosis cannot be over-emphasized. Machine learning helps to structure, normalize, and analyze health data, so healthcare and life science organizations can use it to improve their decisions. This collapses time and transforms health data in a short time, which in turn provides early intervention or diagnosis (Jiang et al., 2017).

As mentioned above, early detection plays a key role in improving patient outcomes, as cervical cancer is highly treatable when identified at precancerous or early invasive stages.

However, many patients in under-served regions with advanced disease are often ill-treated, resulting in poorer prognoses and increased complexity of treatment. Addressing these challenges requires not only improvements in healthcare delivery, but also the development of computational approaches that can support early detection and risk stratification using limited and heterogeneous patient data.

1.6 Limitations of Current Screening and Diagnostic Approaches

Conventional cervical cancer screening methods, including the Papanicolaou (Pap) test and human papillomavirus (HPV) testing, have played a central role in reducing disease incidence. Although effective in many clinical settings, these techniques face several limitations that hinder their widespread adoption and consistent performance. Pap smear interpretation is subject to inter-observer and intra-observer variability, requiring experienced cytotechnologists and pathologists to ensure diagnostic accuracy (Nayar and Wilbur, 2015; Denny et al., 2017). HPV testing, although more sensitive, may suffer from lower specificity and limited accessibility in low-resource settings.

In addition to technical limitations, logistical and socioeconomic factors further restrict the reach of traditional screening programs. Many regions lack the infrastructure required for laboratory-based testing, follow-up procedures, and longitudinal patient follow-up. Delays in result reporting and loss to follow-up remain persistent challenges, particularly in rural and underserved populations.

These limitations have motivated the exploration of automated and semi-automated diagnostic tools capable of helping clinicians interpret clinical data more efficiently and consistently. Artificial intelligence and machine learning techniques offer the potential to com-

plement existing screening methods by providing decision-support systems that are scalable, reproducible, and less dependent on subjective human interpretation.

1.7 Research Objectives and Contributions

The primary objective of this dissertation is to develop and evaluate data-driven computational classification models under data-scarce, imbalanced, and heterogeneous clinical conditions. This work delivers the following contributions:

- **Systematic evaluation of traditional machine learning models (Chapter 4)** for classification under limited and noisy data conditions. This establishes rigorous performance baselines and clarifies the limitations of classical ML in low-resource clinical environments and tabular medical classification.

- **Comprehensive analysis of data augmentation and synthetic data generation techniques (Chapter 4)** including SMOTE, GAN-based synthesis, diffusion models, and LLM-generated tabular data. This contribution demonstrates how generative methods can mitigate class imbalance and improve robustness, advancing augmentation strategies for small, heterogeneous datasets.

- **A novel organized tabular-to-text transformation framework (Chapter 5)** that reformulates structured clinical records into natural-language representations, enabling the use of modern language models for classification. This expands methodological options for handling sparse tabular data and contributes a new technique for bridging structured and unstructured representations.

- **Investigation of Chain-of-Thought, zero-shot, few-shot learning paradigms (Chapter 5)** using large pre-trained language models. This contribution shows that

LLMs can perform competitively with minimal labeled data, offering a scalable approach for medical domains where annotation is costly or infeasible.

- **An extended large-scale clinical taxonomy case study (Chapter 6)** that generalizes the proposed frameworks to more than 165,000 clinical narratives. This chapter demonstrates how unsupervised semantic analysis, latent topic projection, and transformer-based classification can be combined to construct a high-quality *Silver Standard* labeling system. Provides empirical evidence of the scalability, adaptability, and potential domain transferability of the dissertation's methods.

1.8 Organization of the Dissertation

The remainder of this dissertation is organized as follows. Chapter 2 reviews related work in cervical cancer diagnostics, machine learning, data augmentation, and language models. Chapter 3 discusses the key research challenges addressed in this work. Chapters 4 and 5 present the primary studies on machine learning with augmented data and language model-based classification, respectively. Chapter 6 expands the scope of the dissertation by applying similar computational principles including semantic modeling, topic projection, and transformer based classification to a large corpus of more than 165,000 clinical narratives, demonstrating that the methods explored in earlier chapters can scale to broader clinical text analysis tasks. Finally, Chapter 7 synthesizes the findings of the dissertation, draws conclusions from the experimental results, discusses their implications, and identifies directions for future research.

Chapter 2

Related Work

The pursuit of accurate, accessible, and early diagnostic tools for medical data represents a critical frontier in computational oncology and global public health. This chapter synthesizes and critically examines the existing literature in three interconnected domains that underpin this thesis: (i) the evolution of machine learning (ML) models for medical diagnosis and cervical cancer prediction as a case study, (ii) the paramount role of language and communication in healthcare dissemination, and (iii) contemporary strategies for mitigating data scarcity in medical ML, including data augmentation and synthetic data generation. Although significant advancements have been made individually in these fields, a conspicuous gap exists at their intersection: the development of linguistically and culturally contextualized predictive systems that are robust to the data limitations endemic to medical research. This review establishes the foundational knowledge, identifies methodological trends and shortcomings, and ultimately justifies the novel integrative approach proposed in this dissertation, which seeks to harmonize predictive analytics with linguistic accessibility through advanced data synthesis techniques.

2.1 Machine Learning for Medical Diagnosis: An Evolving Landscape

The application of machine learning to medical diagnostics has progressed from utilizing basic classifiers on structured clinical data to employing sophisticated deep learning architectures on multimodal data sources.

2.1.1 Traditional Classifiers and Data Imbalance Challenges

Early and prevalent approaches have used ensemble methods and traditional classifiers in risk-factor datasets. Studies by Abdoh et al. (2018) and Choudhury et al. (2018) exemplify this phase, employing Random Forests alongside techniques such as synthetic minority oversampling technique (SMOTE) and recursive feature elimination to address class imbalance and dimensionality. Kuruvilla and Jayanthi (2023) extended this by integrating ensemble methods (XGBoost, AdaBoost, Random Forest) optimized with bio-inspired algorithms like Firefly, reporting high accuracy. A critical analysis reveals that while these studies report accuracies ranging from 80% to 95%, they often operate on constrained, homogeneous datasets and their real-world clinical generalizability, particularly in diverse linguistic and socioeconomic settings, remains inadequately tested.

2.1.2 Deep Learning and Image-Based Diagnostics

A parallel and transformative strand of research focuses on image-based detection. Hu et al. (2019) demonstrated the potential of deep learning algorithms trained in longitudinal cohorts of digitized cervical images. This shift towards deep learning for Pap smear and colposcopy image analysis (William et al., 2018) promises automated, high-throughput

screening. However, these models are notoriously data-hungry, computationally intensive, and their "black-box" nature poses challenges to clinical interpretability and trust.

2.1.3 Towards Robust and Integrated Frameworks

Recognizing common methodological pitfalls, recent work has proposed more holistic frameworks. Lu et al. (2020), for example, introduced a system that incorporates voting strategies and data correction modules to improve robustness. This evolution signifies a move from isolated model optimization to pipeline-aware solutions that address data quality and integration. Despite these advances, the literature predominantly treats the predictive task in isolation from the crucial subsequent step of patient-facing communication and education. This dissertation focuses primarily on improving methodological robustness within the predictive pipeline; although not directly addressing patient-facing communication, it highlights this gap and discusses the implications and limitations of existing approaches with respect to this downstream need.

2.2 The Critical Role of Language in Healthcare Communication

Effective communication is not only an ancillary concern in healthcare, but a determinant of outcomes. Linguistic theory, dating to Sapir (1921), establishes language as the fundamental framework for thought and social identity. Bamberg (1997) further argues that language constructs reality, influencing how individuals perceive health, risk, and illness. The tripartite functions of language as a medium of communication, a reflector of culture and personality, and a vehicle for cultural transmission (Krech et al., 1962; Fatima et al., 2024) are acutely relevant in medical contexts.

2.2.1 Language as a Social Determinant of Health

Communicating complex health information, such as cervical cancer risks and prevention strategies, in a patient's native or preferred language is imperative for comprehension, informed consent, and adherence to medical advice (Omowhara et al., 2022). The failure to do so can exacerbate health disparities, particularly in developing regions and among marginalized populations where cervical cancer burden is highest. Thus, predictive models achieve limited public health impact if their outputs are not translatable into actionable, linguistically tailored awareness.

2.2.2 The Gap Between Prediction and Communication

Current ML research for cervical cancer, as reviewed in Section 2.1, largely neglects this communicative dimension. In this dissertation, we review both *risk estimation* and *awareness creation* tailored to the linguistic context of the end-user. This necessitates a bridge between structured data-driven prediction and the unstructured domain of natural language generation, a challenge addressed through the methodologies discussed in the following sections.

2.3 Overcoming Data Scarcity: Augmentation in Medical ML

The performance of ML models, especially deep learning, is dependent on large, high-quality datasets. Medical research often faces severe data limitations due to privacy regulations, rare conditions, and costly annotation processes. This section reviews technical strategies to overcome these barriers.

2.3.1 Data Augmentation: Enhancing Real Data

Data augmentation involves applying label-preserving transformations to existing data to increase volume and variability. In medical imaging, techniques such as elastic deformations (Shorten and Khoshgoftaar, 2019) and geometric transformations have proven effective for tasks including colposcopy image analysis (Pereira et al., 2020). For tabular and textual data, analogous techniques include feature noise injection, synthetic sampling (e.g., SMOTE), and textual paraphrasing. Augmentation acts as a regularizer, reducing overfitting and improving model generalization, though it is inherently limited by the diversity present in the original dataset. Chapter 4 presents the primary studies on machine learning with augmented data.

2.3.2 Synthetic Data Generation: Creating Novel Data Instances

A more powerful paradigm involves generating entirely new, realistic data instances. Generative adversarial networks (GANs) and their variants, such as conditional GANs (cGANs), have shown remarkable success in creating synthetic medical images. Zhao et al. (2021) Demonstrated this for cervical cell images, preserving critical pathological features. In the textual domain, generative models can synthesize patient records or descriptive text. Synthetic data can address class imbalance more fundamentally and create datasets for scenarios where real data is prohibitively scarce or sensitive. Chapter 4 presents the primary studies on machine learning with synthetic data.

2.3.3 Integrated and Hybrid Approaches

The integration of augmentation and synthetic generation presents a comprehensive strategy. Combining these methods can create richly diverse and extensive training corpora, leading to more robust and generalizable models. However, key challenges persist, including evaluating

the fidelity and clinical validity of synthetic data, avoiding the propagation of biases, and navigating the ethical implications of using non-real patient data in model development.

2.4 Artificial Intelligence and Language Models for Medical Diagnostics

The rise of large language models (LLMs) and transformer-based architectures has revolutionized natural language processing (NLP), offering new tools for medical text analysis.

2.4.1 Transformer Models and Their Biomedical Adaptations

Models such as Bidirectional Encoder Representations from Transformers (BERT) (Vaswani et al., 2017) introduced deep bidirectional context understanding. For computational efficiency, distilled versions such as DistilBERT (Sanh et al., 2019b) and parameter-efficient models such as A Lite Bert (ALBERT) (Lan et al., 2019b) were developed. Crucially, domain-specific pre-training has led to powerful biomedical variants such as BioBERT (Lee et al., 2020) (trained on biomedical literature) and PubMedBERT (Gu et al., 2020a) (trained on PubMed abstracts), which significantly outperform general models on tasks like named entity recognition and relation extraction from clinical text.

2.4.2 Data Augmentation and Few-Shot Learning in Medical NLP

In medical NLP, data augmentation techniques such as back-translation, synonym replacement, and generative AI for synthetic text creation (Shorten and Khoshgoftaar, 2019), are vital for overcoming data paucity. Furthermore, few-shot and zero-shot learning paradigms (Brown et al., 2020) enable models to make reliable predictions with minimal labeled examples by leveraging knowledge encoded during pre-training. This is particularly valuable

for rare diseases or low-resource languages where annotated clinical corpora are limited. Chapter 5 presents the primary studies on machine learning with language model-based classification.

2.4.3 Gap in Application to Medical Classification

While transformer models have been extensively applied to clinical note analysis and biomedical literature mining, their application to medical diagnosis, especially cervical cancer classification, has been narrow. Existing AI research in this domain remains dominated by image-based CNNs (William et al., 2018). There is a nascent opportunity to repurpose these advanced NLP models for classification tasks derived from *textual representations* of patient data, rather than only images. This approach can encapsulate a wider range of risk factors (demographic, behavioral, clinical history) in a format that language models are inherently designed to process.

2.5 Synthesis and Identified Research Gap

This review delineates three robust yet largely siloed bodies of work:

1. **Predictive ML Models**: Focused on accuracy but often clinically and linguistically disembodied.

2. **Health Communication Theory**: Emphasizes linguistic and cultural congruence but is rarely integrated with predictive analytics.

3. **Data-Centric AI**: Provides tools (augmentation, synthetic data, LLMs) to overcome data limitations and process unstructured text.

The core research gap, therefore, is the absence of an integrated and operational approach to clinical risk modeling that addresses three critical challenges:

(a) robust data synthesis mechanisms (including augmentation and generation) to overcome the scarcity and imbalance of labeled medical data while preserving clinical relationships (Chapter 4); (b) effective, domain-adapted language modeling approaches capable of transforming and processing heterogeneous patient information into contextually rich representations for accurate classification (Chapter 5); and (c) structured and interpretable output representations that enable clinically meaningful, multi-label predictions and support downstream generation of accessible health communication for risk awareness (Chapter 6).

2.6 Analysis of Clinical Narratives

Clinical documentation represents a rich but underutilized resource for patient stratification and cohort discovery. Chapter 6 represents a comprehensive methodology for unsupervised clustering of clinical narratives using a hybrid approach that combines deep semantic embeddings from transformer models with structured clinical feature extraction. We evaluated our method on a corpus of 50,000 patient summaries from PubMed Central (PMC), comparing pure BERT-based clustering against our hybrid approach. The results demonstrate that while pure semantic clustering yields modest quantitative metrics, the hybrid approach significantly improves both cluster cohesion and clinical interpretability. Our findings highlight the importance of integrating domain knowledge with deep learning for meaningful patient stratification in electronic health records.

2.7 Clinical Natural Language Processing

Clinical NLP has evolved significantly over the past two decades. Early systems relied on rule-based approaches and dictionary lookups. The advent of statistical methods and machine learning brought improvements in named entity recognition and relation extraction. More recently, deep learning approaches have achieved state-of-the-art performance in various clinical NLP tasks (Fu et al., 2020).

Transformer models, particularly BERT and its clinical variants, have shown remarkable success in capturing semantic relationships in medical text. ClinicalBERT and BioClinicalBERT are pre-trained on large corpora of clinical notes and biomedical literature, making them particularly suited for clinical text analysis (Diaz Ochoa et al., 2025).

2.7.1 Clinical Text Mining Landscape

The digitization of healthcare has generated unprecedented volumes of unstructured clinical text, including physician notes, discharge summaries, radiology reports, and pathology findings. These documents contain rich phenotypic information that remains largely untapped for research and clinical decision support (Holmes et al., 2021). Traditional approaches to patient stratification are based on structured data elements such as diagnosis codes, laboratory values, and medication records (Aljohani et al., 2024). However, these structured representations often fail to capture the nuanced clinical context, temporal evolution, and severity information embedded in free-text narratives (Diaz Ochoa et al., 2025).

Clinical text mining faces unique challenges compared to general domain natural language processing (NLP). Medical language exhibits high lexical variation, extensive use of abbreviations and acronyms, complex negation patterns, and domain-specific syntactic structures (Friedman et al., 2004). Furthermore, clinical narratives often contain multiple co-occurring

conditions, making patient-level classification particularly challenging (Koleck et al., 2019).

2.7.2 Text Clustering Methods

Text clustering algorithms can be broadly categorized into several approaches:

- Traditional Vector Space Models: Early text clustering methods employed bag-of-words representations with dimensionality reduction techniques such as latent semantic analysis (LSA) (Landauer et al., 1998) and non-negative matrix factorization (NMF) (Lee and Seung, 1999). These methods, while interpretable, often fail to capture semantic relationships between terms.

- Topic Modeling: Probabilistic topic models such as latent Dirichlet allocation (LDA) (Blei et al., 2003) and its extensions have been widely used for document clustering. Clinical applications of topic modeling include identifying themes in clinical notes (Arnold et al., 2010) and discovering comorbidities (Perotte et al., 2014).

- Deep Learning Approaches: Deep learning methods for text clustering include autoencoder based approaches (Vincent et al., 2010), deep embedded clustering (Xie et al., 2016), and transformer-based methods (Reimers and Gurevych, 2019). These approaches can learn hierarchical representations that capture complex semantic relationships.

2.7.3 Patient Stratification in Precision Medicine

Patient stratification, the process of dividing heterogeneous patient populations into clinically significant subgroups, is fundamental to precision medicine (Collins et al., 2017). Traditional stratification approaches typically rely on supervised learning with labeled outcomes or phenotypes defined by experts. However, unsupervised methods offer the potential to

discover new patient subgroups without predefined labels, potentially revealing previously unrecognized disease subtypes or treatment response patterns (Lopez et al., 2018).

Unsupervised clustering of clinical text presents particular challenges due to High dimensionality of text representations, the curse of dimensionality in clustering algorithms, the need for clinically interpretable results, handling of medical terminology and abbreviations, and integration of temporal information.

2.7.4 Hybrid Approaches in Clinical Analytics

Several studies have explored hybrid approaches that combine structured and unstructured data in clinical analytics. Wang et al. (2018b) combined clinical notes with structured EHR data for the prediction of mortality, while Li et al. (2019) integrated text and time-series data for the detection of sepsis. However, few studies have explored hybrid approaches specifically for unsupervised patient stratification from clinical text.

2.8 Chapter Summary

This chapter has established the necessary theoretical and technical foundation for the present research. It critically analyzed the progression of cervical cancer prediction models, underscored the non-negotiable importance of linguistically tailored communication, and reviewed state-of-the-art methods for data enrichment and language-based classification. The synthesis of these strands reveals a compelling opportunity for innovation. The subsequent chapters of this thesis will detail a novel methodology that addresses this identified gap by transforming structured patient data into textual descriptions, employing generative techniques for data augmentation, and fine-tuning a suite of compact language models (e.g., BERT, PubMedBERT, LLaMA) for accurate classification. Subsequently, we dived into the

burgeoning field of clinical text mining, highlighting its unique challenges and the rich, untapped phenotypic information within unstructured healthcare narratives. We reviewed various text clustering methods, from traditional vector space models to modern deep learning approaches, recognizing their potential for discovering hidden patient subgroups. Furthermore, we examine the landscape of patient stratification in precision medicine, emphasizing the power of unsupervised methods to reveal novel disease subtypes and treatment response patterns. The exploration of hybrid approaches in clinical analytics demonstrated the growing interest in the integration of structured and unstructured data, though a specific focus on unsupervised patient stratification from clinical text remains an area ripe for exploration. This approach not only aims to advance predictive performance, but also to bridge the critical last mile between algorithmic prediction and patient understanding, contributing to more equitable and effective cervical cancer prevention strategies.

Chapter 3

Research Challenges and Problem Formulation

The development of computational models for medical diagnosis and classification represents a critical intersection of machine learning innovation and urgent public health need. As established in the literature review, while significant progress has been made in predictive modeling, data augmentation, and natural language processing for medical applications, these advancements remain largely siloed. This chapter synthesizes the identified limitations from prior research and formalizes the specific interconnected challenges that this dissertation addresses. The challenges stem from three core tensions: (1) the conflict between the data-hungry nature of advanced models and the severe data constraints of real-world medical contexts, particularly in low-resource settings; (2) the gap between achieving algorithmic accuracy and ensuring clinical utility through interpretable, actionable outputs; and (3) the methodological disconnect between structured data analysis and unstructured language-based reasoning in patient care pathways. This chapter articulates these challenges as a series of research questions that guide the methodological development and experimental

validation in subsequent chapters.

3.1 Challenge 1: The Data Scarcity-Imbalance in Medical Machine Learning

This section introduces a central difficulty in medical machine learning: the simultaneous presence of limited data availability and class imbalance.

3.1.1 Nature of the Challenge

Medical datasets, such as the University of California, Irvine (UCI) cervical cancer dataset (Fernandes et al., 2017) utilized in this work, are characterized by a fundamental paradox: they are simultaneously *too small* to train complex modern models and *too imbalanced* for reliable clinical prediction. The dataset from 'Hospital Universitario de Caracas' ($n = 858$), while valuable, exemplifies this issue. Its limited size restricts the ability to train deep learning architectures without severe overfitting. Further, class imbalance between positive and negative cases biases models toward the majority class, potentially masking early-stage indicators crucial for screening.

3.1.2 Specific Limitations of Existing Solutions

Current approaches to this paradox, as reviewed in Chapter 2, offer partial solutions with significant drawbacks:

- **Oversampling Techniques (e.g., SMOTE):** While effective for simple classifiers, SMOTE generates synthetic samples through linear interpolation in feature space, which can create unrealistic or noisy data points in complex, high-dimensional medical

feature spaces, potentially degrading model generalizability.

- **Traditional Data Augmentation:** Methods such as geometric transformations are ill-suited for structured, heterogeneous tabular data comprising categorical (habits, history) and numerical (age, counts) variables.

- **Generative Models in Medical Contexts:** The application of GANs and diffusion models to tabular medical data remains nascent. Key challenges include preserving complex statistical dependencies between clinical variables (e.g., the relationship between smoking habits, HPV status, and biopsy results) and ensuring the clinical plausibility of generated synthetic patient records to avoid propagating spurious correlations.

3.1.3 Formal Research Questions

This challenge leads to the following research questions:

1. How do advanced generative models (e.g., GANs, forest diffusion, and LLM-based synthesis) improve classifier performance on the imbalanced UCI cervical cancer dataset?

2. Can generative models produce synthetic tabular patient data that preserves clinically meaningful relationships necessary for robust model training, as evidenced by data distribution analysis and downstream classification performance?

3. What is the optimal integration strategy for combining original and synthetically augmented data to maximize model generalization and minimize overfitting in a low-data regime?

3.2 Challenge 2: The Modality Gap: Bridging Structured Data and Contextual Understanding

This section introduces the modality gap between structured medical data and the richer contextual reasoning used in real clinical decision-making. It explains how current machine learning models fail to capture the narrative and semantic relationships embedded in patient data.

3.2.1 Nature of the Challenge

Clinical decision-making is inherently multimodal, integrating structured data (lab values, risk factors) with unstructured contextual knowledge (patient history, clinical notes). Traditional ML models excel at processing the former but fail to capture the latter. This creates a *modality gap*: a patient's risk profile is more than the sum of discrete features; it is a narrative. Existing models that use the UCI dataset treat variables as independent or shallowly correlated inputs, missing the holistic story of a patient's health status.

3.2.2 Limitation of Tabular-Feature-Only Approaches

Models like random forests or gradient boosting, while powerful, process features in isolation from the rich semantic relationships that a clinician would consider. For example, the combination of "number of sexual partners," "smoking years," and "first sexual intercourse age" forms a risk narrative that is more informative than each feature independently. Traditional models lack a native mechanism to learn these high-order, semantic interactions effectively.

3.2.3 Formal Research Questions

This challenge leads to the following research questions:

1. Can reformulating structured tabular patient records into natural language descriptions (e.g., "A 35-year-old non-smoker with 3 sexual partners...") enable language models to capture richer, contextual relationships for improved classification?

2. Do domain-specific models (PubMedBERT, BioBERT), pretrained on biomedical corpora, outperform general-purpose models when applied to textual patient representations?

3. Does a text-based representation of patient data provide a robust framework for modeling contextual relationships among clinical risk factors?

3.3 Challenge 3: The Deployment Dilemma: Accuracy vs. Accessibility in Real-World Settings

This section addresses the practical challenge of deploying medical machine learning models in real-world clinical environments, particularly in resource-constrained settings. It highlights the trade-off between achieving high predictive accuracy and maintaining computational efficiency and accessibility.

3.3.1 Nature of the Challenge

The utility of a model is determined not only by its accuracy in validation but also by its feasibility in the resource-constrained environments where cervical cancer burden is greatest. There exists a fundamental tension between model complexity and deployability. Large,

state-of-the-art models (e.g., deep neural networks, massive LLMs) may achieve marginally higher accuracy but require computational resources, inference time, and expertise that are unavailable in low- and middle-income country (LMIC) clinics.

3.3.2 The Inefficiency of State-of-the-Art Models

As identified in the literature, models achieving $90-95\%$ accuracy often rely on ensembles or deep architectures that are impractical for real-time, point-of-care screening. Furthermore, the push for higher accuracy on benchmark datasets like UCI can lead to over-engineering that fails to translate to diverse, real-world populations. The challenge is to develop a framework that prioritizes efficient sufficiency achieving robust, and clinically acceptable performance with minimal computational footprint.

3.3.3 Formal Research Questions

This challenge leads to the following research questions:

1. Can data-augmented machine learning models and domain-specific language models provide an effective balance between predictive performance and computational efficiency?

2. To what extent can zero-shot and few-shot learning with pre-trained LLMs provide a viable, low-data alternative to fully supervised training, thereby reducing the labeled data burden for deployment in new clinical settings?

3. How effectively can the proposed framework operate in a low-data medical environment characterized by limited labeled samples and class imbalance?

3.4 Challenge 4: The Interpretability-Actionability Disconnect

This section addresses the gap between model predictions and their practical usefulness in clinical settings. The section discusses the limitations of current explainability methods and explores whether language-based modeling can provide more intuitive, clinically meaningful explanations.

3.4.1 Nature of the Challenge

Even an accurate model has limited clinical impact if its output is not interpretable and actionable for healthcare providers. A binary risk score (e.g., "high risk") is insufficient; clinicians require insights into *why* a patient is flagged to inform subsequent steps (e.g., type of follow-up test, patient counseling). This is intrinsically linked to the role of language in healthcare communication, as discussed in Section 2. Current ML models for cervical cancer prediction are largely black boxes, offering predictions without explanations.

3.4.2 The Unfulfilled Promise of Explainable AI (XAI)

Although techniques like SHapley Additive exPlanations (SHAP) or Local Interpretable Model-agnostic Explanations (LIME) (Salih et al., 2025) can provide post-hoc feature importance for tabular models, they often generate abstract, statistical explanations (e.g., "Smoking years contributed +0.2 to the risk score") that are not easily translatable into clinical dialogue or patient-facing communication.

3.4.3 Formal Research Question

This challenge, while not the primary focus of the experimental chapters, underpins the motivation for the text-based approach and leads to a forward-looking research question:

1. Does the proposed text-based modeling framework, by its very architecture, offer a more natural pathway towards generating interpretable, textual explanations of model predictions (e.g., "The model indicates high risk primarily due to the patient's long-term smoking habit combined with a history of STDs") compared to feature-attribution methods applied to tabular models?

3.5 Synthesis: The Integrated Research Problem

The challenges outlined above are not independent; they form a interconnected web of constraints that define the core research problem of this dissertation.

The central hypothesis of this work is that an integrated framework leveraging generative data augmentation and language-based representation learning can simultaneously address these challenges. Specifically:

- By using advanced synthetic data generation, we can mitigate data scarcity and class imbalance while improving classifier performance.
- By transforming data into text and applying domain-specific language models, we can capture richer clinical context and semantic relationships.
- By evaluating efficient machine learning and language-model approaches, we can develop solutions suitable for deployment in resource-constrained environments.
- This integrated framework also creates opportunities for more interpretable and clinically meaningful model outputs.

3.6 Research Objectives

The primary research objective of this dissertation is to develop and rigorously evaluate machine learning pipelines for cervical cancer classification that effectively address the challenges of data scarcity, modality gaps, deployment constraints, and interpretability. This involves assessing various data augmentation techniques including traditional oversampling, generative adversarial networks, diffusion models, and large language model based synthesis to enhance model robustness in low-resource settings.

The research aims to reformulate structured clinical data in natural language and use domain-specific and general-purpose language models to improve classification performance and contextual understanding. The study also extends to the integration of real clinical narratives and multimodal patient information to enrich model inputs, while systematically analyzing the robustness, interpretability, and clinical relevance of all approaches to ensure practical applicability in the real-world of healthcare.

In addition, this research makes several key contributions such as: methodological innovation: We propose a hybrid clustering framework that integrates semantic embeddings from domain-specific BERT models with structured clinical feature extraction. Comprehensive evaluation: we evaluate our approach using multiple clustering metrics and, importantly, clinical interpretability measures. Practical implementation: we provide a complete, reproducible pipeline for clinical text clustering, including preprocessing, feature engineering, clustering, and visualization. Clinical insight generation: we demonstrate how our method can generate clinically meaningful patient subgroups that align with medical specialties and disease categories.

3.7 Chapter Summary and Link to Methodology

This chapter has systematically dissected the landscape of cervical cancer classification to identify four principal research challenges: the Data Scarcity-Imbalance Paradox, the modality gap, the deployment dilemma, and the interpretability-actionability disconnect. These challenges have been formalized into ten specific research questions mentioned earlier in this chapter that provide clear, testable objectives for this dissertation.

The following sections directly answer these questions. Chapter 5 details the proposed two-part methodology: (1) a comprehensive evaluation of traditional ML with advanced data augmentation, (2) the novel framework for text-based classification using language models, and (3) a comprehensive methodology for unsupervised analysis of clinical narratives. Chapters 4, 5, and 6 present the experimental setup and results, structured to provide empirical answers to each research question posed here. Through this structured approach, the dissertation advances the field not by pursuing marginal accuracy gains in isolation, but by developing a more holistic, robust, and deployable paradigm for computational cervical cancer risk assessment.

Chapter 4

Machine Learning Classification with Synthetic and Augmented Data

This chapter discusses and describes the proposed methodology for classifying the risk of cervical cancer. The methodology includes a detailed overview of the dataset, preprocessing steps, and the generation of synthetic data using various approaches including including generative adversarial networks (GANs), conditional tabular generative adversarial networks (CTGANs), forest diffusion, and large language models (LLMs). This study is published in IEEE International Conference on Prognostics and Health Management (ICPHM) (Gheibi and Hougen, 2025a).

4.1 Dataset

The Cervical Cancer Risk Factors (CCRF) dataset (Fernandes et al., 2017) used in this study comes from the University of California, Irvine (UCI) database. It encompasses a collection of patient histories, medical practices, procedures, and demographic data across 858 cases,

each comprising 36 features in which the first 32 features in the dataset capture personal and medical history, while the last 4 columns: *Hinselmann*, *Schiller*, *Citology*, and *Biopsy* represent whether a healthcare provider recommended further cervical cancer screening tests. In particular, the dataset includes several missing attributes due to incomplete records. These omissions were deliberately maintained to avoid addressing privacy concerns, reflecting the prevalent challenges of privacy and security in healthcare record management systems. When deciding whether to generate synthetic data from processed or raw data, several factors must be taken into account:

Data Quality and Characteristics:

- Raw Data: Using raw data to generate synthetic versions can ensure that all inherent variations and noise present in the original dataset are retained. This approach is beneficial for scenarios that require resilience to real-world conditions.

- Processed Data: For raw data that include errors, irrelevancies, or excessive noise, preprocessing before synthetic data generation can lead to enhanced quality in synthetic datasets. This approach is particularly valuable when the end use requires clean, well-structured data.

Application Requirements:

- For use cases that require maintaining the original data complexity and realism, generating synthetic data from raw data is preferable (Petrakos et al., 2025).

- In scenarios where the focus is on specific attributes or cleaner datasets (such as in certain machine learning applications), generating from processed data might be more beneficial.

Privacy Concerns:

- Processed data can be advantageous for synthetic data generation when privacy is a priority, as sensitive information can be altered or removed during preprocessing.

Training and Model Performance:

- Synthetic data derived from processed data can be customized to emphasize relevant features, potentially improving the training and performance of analytical models.
- Alternatively, synthetic data from raw data can promote greater generalizability and prevent overfitting in models by reflecting the true complexity of original datasets.

The choice between raw and processed data for generating synthetic datasets should be guided by the specific goals and requirements of a project, including the desired features of the synthetic data and the constraints of the original data.

4.2 Preprocessing

This section outlines the data preparation pipeline applied to the cervical cancer dataset, emphasizing techniques used to ensure data quality, integrity, and suitability for downstream modeling. It covers strategies for handling missing data, feature selection, scaling, and dataset restructuring. The section also motivates the importance of careful preprocessing in small, imbalanced datasets and transitions into advanced data augmentation approaches (GANs, CTGAN, diffusion models, and language-model-based generators) used to address these limitations.

Data preprocessing is essential for handling missing values, normalizing numerical attributes, and removing redundant features. The key preprocessing steps include the following.

- **Retaining Duplicates:** The original dataset contains 858 rows, however several entries are exact duplicates across all 32 features. Duplicate records were retained in the final dataset due to the absence of unique patient identifiers. Given that many features consist of low-cardinality integers or binary reals (e.g., $0.0, 1.0$), identical feature vectors can represent distinct clinical instances rather than redundant entries. Removing these records without evidence of a non-random origin risks altering the underlying data distribution. This shift is particularly problematic for our imputation strategy, which relies on preserving the original distribution to maintain integrity. Consequently, the analysis was conducted using the complete dataset to ensure a representative sample and to avoid biasing the results of the imputation and subsequent algorithms.

- **Handling Missing Data:** Missing values can introduce bias and reduce model reliability. Techniques such as multiple imputation and mean/mode substitution are commonly used (Pham et al., 2024).

- **Feature Selection:** Irrelevant or redundant features can negatively impact model performance. Methods like recursive feature elimination (RFE) and correlation analysis help identify the most important predictors (Bolón-Canedo and Alonso-Betanzos, 2015).

- **Normalization and Standardization:** Generally, machine learning models perform better when numerical features are scaled to a uniform range, ensuring a fair weight distribution across attributes.

- **Structural Missingness and Imputation Strategy:** In the preprocessing phase, handling missing values is crucial to ensure the integrity and accuracy of the analysis. Various techniques, such as imputation, where missing values are replaced with statistical estimates (e.g. mean, median), or deletion, where records with missing values are

entirely removed, can be employed based on the nature of the data and the specific requirements of the study. Proper handling of missing data helps minimize bias and improves the robustness of the model outcomes.

The CCRF dataset exhibits a significant variation in the number of missing values between features, ranging from 7 to 787, as demonstrated by the attribute *STDs: Time since the first diagnosis*, and *STDs: Time since the last diagnosis* as shown in Figure 4.1. These two features were removed due to the number of missing values.

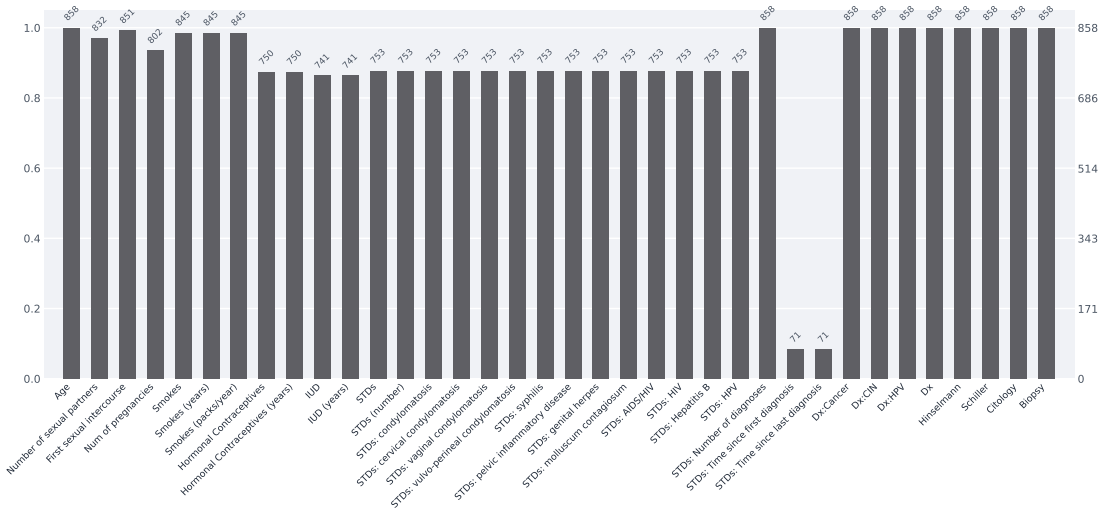

Figure 4.1: Dataset Features and Missing Values

In addressing missing data, a distinction was made between stochastic missingness and structural missingness. Although standard imputation techniques are suitable for data missing at random (MAR) and missing completely at random (MCAR), they may introduce significant noise when applied to features that are missing not at random (MNAR). The missingness categories (MCAR, MAR, MNAR) were determined using the standard definitions from the statistical literature on missing data mechanisms. A feature was classified as MCAR when the absence of a value had no logical dependency on any patient characteristic or on the value of the feature itself. Demographic variables such as age and number of

pregnancies fall into this category.

A feature was classified as MNAR when missingness is directly tied to the value of the feature. This applies to all conditional medical features in the dataset. For example, *STDs: HPV* is missing for patients who have never had an STD. In this case, the missingness is caused by the true value of the feature (zero years since diagnosis), making it MNAR. The same logic applies to all STD related variables and diagnostic test results (*Dx:Cancer*, *Dx:HPV*, *Dx:CIN*), which are only recorded when the corresponding test is performed. For the features with MNAR missingness, we removed those features.

A feature was classified as MAR when missingness could plausibly depend on other observed variables but not on the value of the feature itself. For example, *Smoking years* may be missing when the patient does not smoke or does not recall the duration, but the missingness is not caused by the true value of the variable. For the features with MAR missingness, we used k-nearest neighbors (KNN) imputation. The selected features are: *Smokes*, *Smoking years*, *Hormonal contraceptive use*, and *IUD use*.

Finally, for the features with MCAR missingness, we implemented two methods for imputing missing values: applying the mean of each feature and filling missing values based on the proportional representation of existing values within the features. The selected features are: *Age, Number of sexual partners, First sexual intercourse age, and Number of pregnancies*. For example, the feature *Number of pregnancies* contains 12 distinct values, distributed in Table 4.1. Since this process was done before removing the MNAR records the numbers in Table 4.1 are before removing the records with MNAR.

For example, using mean or median imputation would disproportionately bias the *Number of pregnancies* feature toward 0.0, the most frequent value. Instead, our chosen method filled the 56 missing entries according to the empirical distribution of the existing values, resulting in a more balanced representation. This approach avoids introducing skewed or artificial

Value	Original Count	Original %	Imputed Addition	Final Count	Final %
1.0	270	33.7%	+19	289	33.7%
2.0	240	29.9%	+17	257	30.0%
3.0	139	17.3%	+10	149	17.4%
4.0	74	9.2%	+5	79	9.2%
5.0	35	4.4%	+2	37	4.3%
6.0	18	2.2%	+1	19	2.2%
0.0	16	2.0%	+1	17	2.0%
7.0	6	0.7%	+1*	7	0.8%
8.0	2	0.2%	+0	2	0.2%
10.0	1	0.1%	+0	1	0.1%
11.0	1	0.1%	+0	1	0.1%
Total*	**802**	**100.0%**	**+56**	**858**	**100.0%**

* Final value adjusted by rounding the highest decimal remainder.
** Counts shown before removing MNAR records.

Table 4.1: Feature Values Before and After Filling Missing Values [Number of Pregnancies]

values and helps to maintain the robustness of the dataset for downstream classification tasks. This classification is consistent with established definitions of missing data mechanisms in the statistical literature.

In addition to imputing missing values and eliminating duplicate records, we consolidated the target variables—*Hinselmann*, *Schiller*, *Citology*, and *Biopsy* into a single composite target variable. A record was assigned a positive label when at least one of the four outcomes was positive and a negative label otherwise. This transformation was motivated by the objective of the present study, which is to investigate whether available demographic and clinical risk factors can be used to distinguish between subjects with and without a positive finding among the four diagnostic outcomes. By combining the four target variables into a single binary outcome, the prediction task is simplified to determining whether a subject

may warrant further examination rather than predicting the recommendation of a specific diagnostic test. This transformation altered the structure of the dataset from (858, 36) to (753, 30), which comprises 30 features and one target variable.

Given the small size and class imbalance of the UCI Cervical Cancer Risk Factors dataset, a single train-test split is not sufficient to provide statistically reliable performance estimates. To ensure rigorous evaluation, all models were trained and tested using 10-fold cross-validation.

The dataset was randomly shuffled and partitioned into ten equally sized folds. For each iteration, nine folds (90% of the data) were used for training and the remaining fold (10% percent) was used for testing. This process was repeated ten times so that each fold served as the test set exactly once. For each metric, the mean and standard deviation across the ten folds were computed.

In Mehmood et al. (2021), the authors implemented a feature selection strategy using a random forest model to identify the ten most significant features. These features included *Age*, *Duration of hormonal contraceptive use*, *Age at first sexual intercourse*, *Number of pregnancies*, *Duration of IUD use*, *Years of smoking*, *Packs of cigarettes per year*, *Hormonal contraceptives*, *Time since first STD diagnosis*, and *Time since last STD diagnosis*. Using these selected features, they applied a shallow neural network comprising a single layer with 50 neurons, trained over 10 epochs. In our implementation of the Mehmood et al. (2021) model with our preprocessing, the convergence of the model required 165 iterations, achieving a loss of 0.296. Table 4.2 shows the result for the confusion matrix with all 32 features and the same information for the 10 top ranked features from random forest algorithm.

The evaluation of machine learning classification models usually relies on key metrics such as precision, recall, accuracy, specificity, and the F1-score. These metrics are derived from the four distinct outcomes false positive (FP), false negative (FN), true positive (TP), and true

Dataset	Accuracy	Precision	Recall	Specificity	F1-Score
All Features	**86.23**	**20.00**	**5.00**	**97.28**	**8.00**
Top 10 Features	85.63	0.00	0.00	**97.28**	0.00

Table 4.2: Comparison of MLP Classifier in Mehmood et al. (2021) Results (All Features vs. Top 10 Features)

negative (TN) presented in a confusion matrix. The tables presented herein demonstrate that despite achieving high accuracy, the model exhibits sub-optimal performance in identifying true positives. Although numerous studies highlight the elevated accuracy of their models, this metric is often inflated by the prevalence of true negatives. However, the more critical metric should be the number of true positives, as it significantly enhances the system's value. This chapter prioritizes improving the detection of true positives to develop a robust cancer classification system while simultaneously maintaining a high overall accuracy.

4.2.1 GAN

GANs are a class of artificial intelligence algorithms used in unsupervised machine learning, implemented by a system of two neural networks competing with each other in a zero-sum game framework. This technique was introduced in Goodfellow et al. (2014). The system consists of two models: the generator (G) and the discriminator (D). The generator tries to produce data that are indistinguishable from genuine data, while the discriminator evaluates the authenticity of the data, determining whether each piece of data is real or produced by the generator (Xu et al., 2019). Mathematically, his relationship can be described by the generator G and the discriminator D expressed as

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{\text{data}}}[\log D(x)] + \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z)))].$$

The terms of this equation are defined as follows:

- $\min_G \max_D V(D, G)$: This represents the minimax game between the generator G and the discriminator D. The generator G tries to minimize this function against the discriminator that tries to maximize it.
- $\mathbb{E}_{x \sim p_{\text{data}}}$: The expectation of the real data distribution p_{data}.
- $\log D(x)$: The logarithm of the probability that the discriminator D correctly classifies real data as real.
- $\mathbb{E}_{z \sim p_z}$: The expectation over the distribution of input noise z, typically assumed to be a simple distribution like a Gaussian or uniform distribution.
- $\log(1 - D(G(z)))$: The logarithm of the probability that the discriminator D correctly classifies fake data (produced by the generator G from noise z) as fake.

4.2.2 CTGANs

CTGANs are a specialized extension of the generative adversarial network framework designed to generate synthetic tabular data. This model involves training a generator to produce synthetic data instances and a discriminator to distinguish between real and synthetic data, optimized through a min-max game between two competing neural networks.

After completing the preprocessing steps outlined in Section 4.2, we utilized the CTGAN method on the training portions of our dataset with missing values filled by mean values. The model is configured with a learning rate of 0.02, a max depth of 2 and 100 estimators. The data generation process includes a pre-generation fraction of 2 and training with a batch size of 500, 25 epochs of patience, and 500 total epochs. Importantly, we preserve the integrity of the test set by keeping it unseen and planning to use it exclusively for evaluation. Figure

4.2 shows the comparison between the generated data (fake) and the original data (real) for the first four features of the CCRF dataset. Figures 4.3, 4.4, and 4.5 represent the heatmap of the real, fake and the difference between real and fake dataset features respectively.

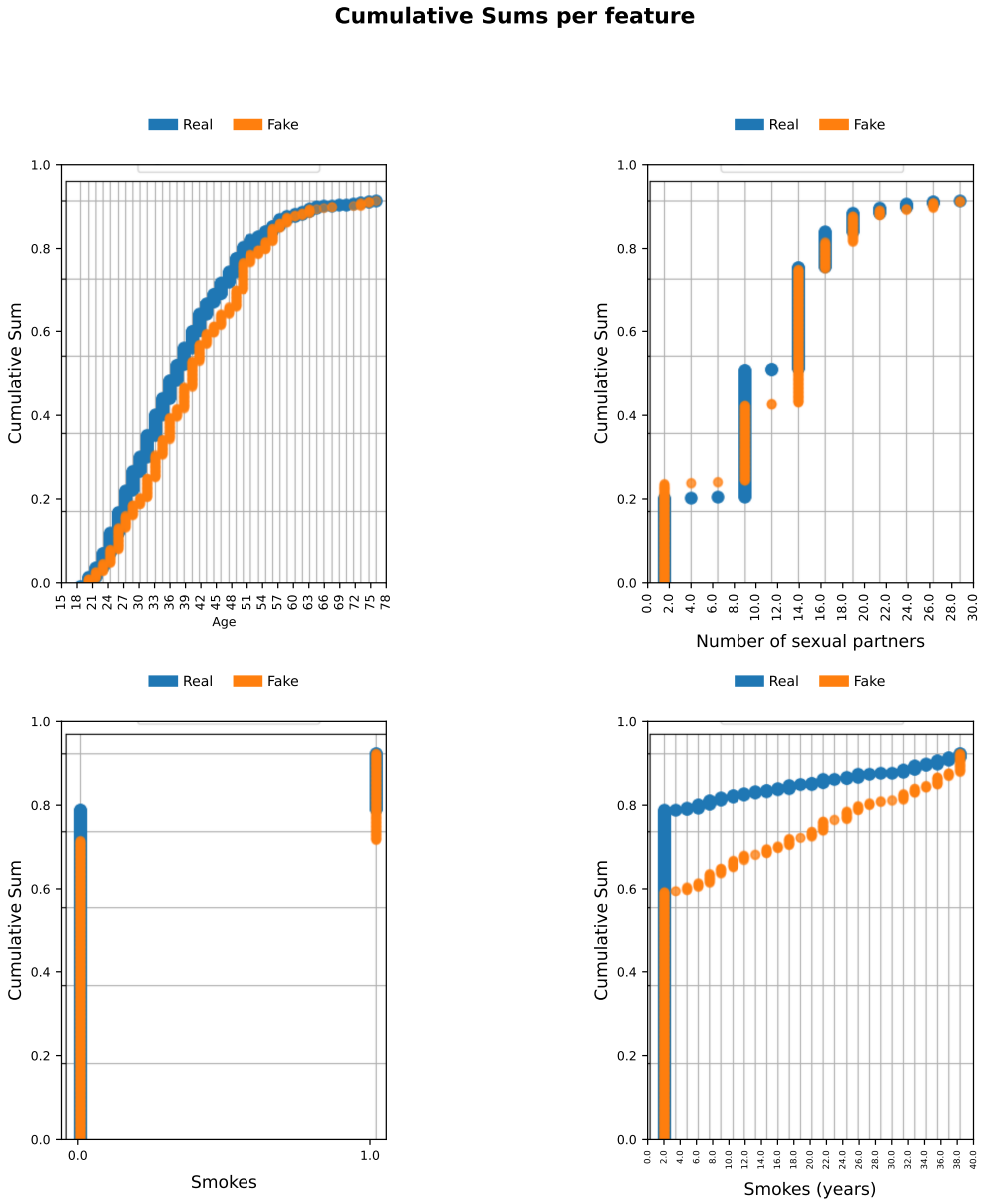

Figure 4.2: Comparison of Real and Fake Data for the First Four Features.

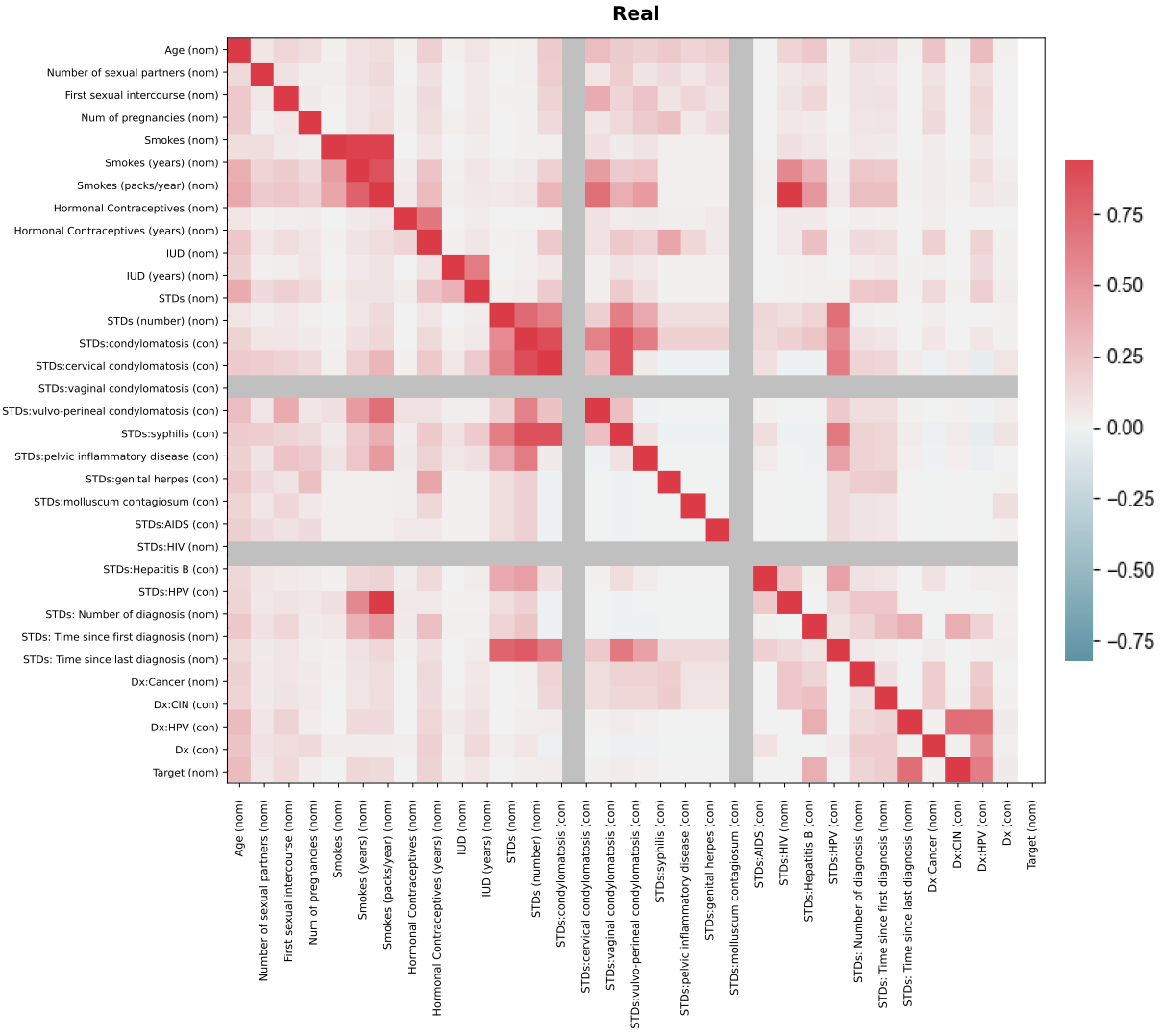

Figure 4.3: Heatmap of the Real Dataset.

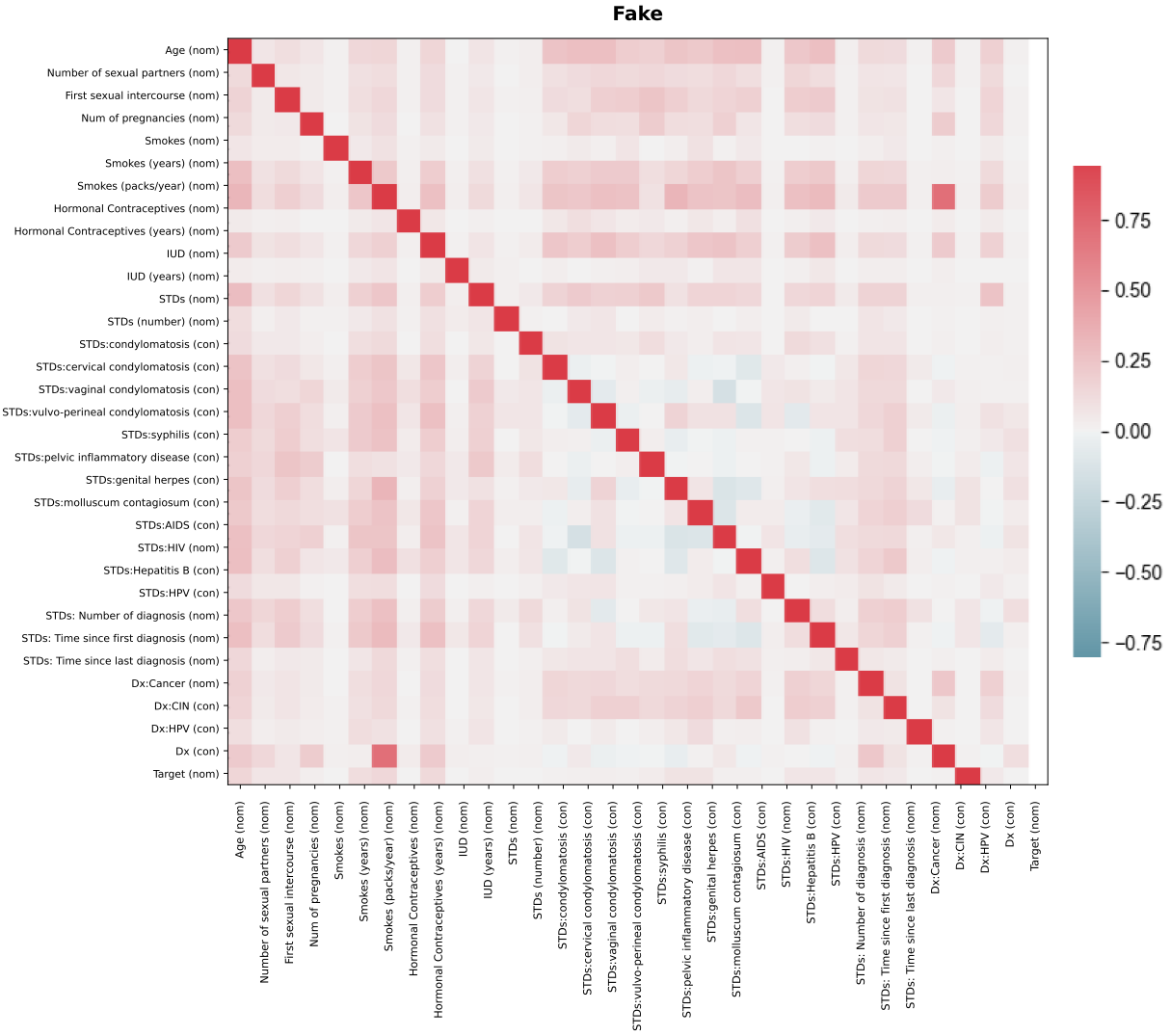

Figure 4.4: Heatmap of the Fake Dataset.

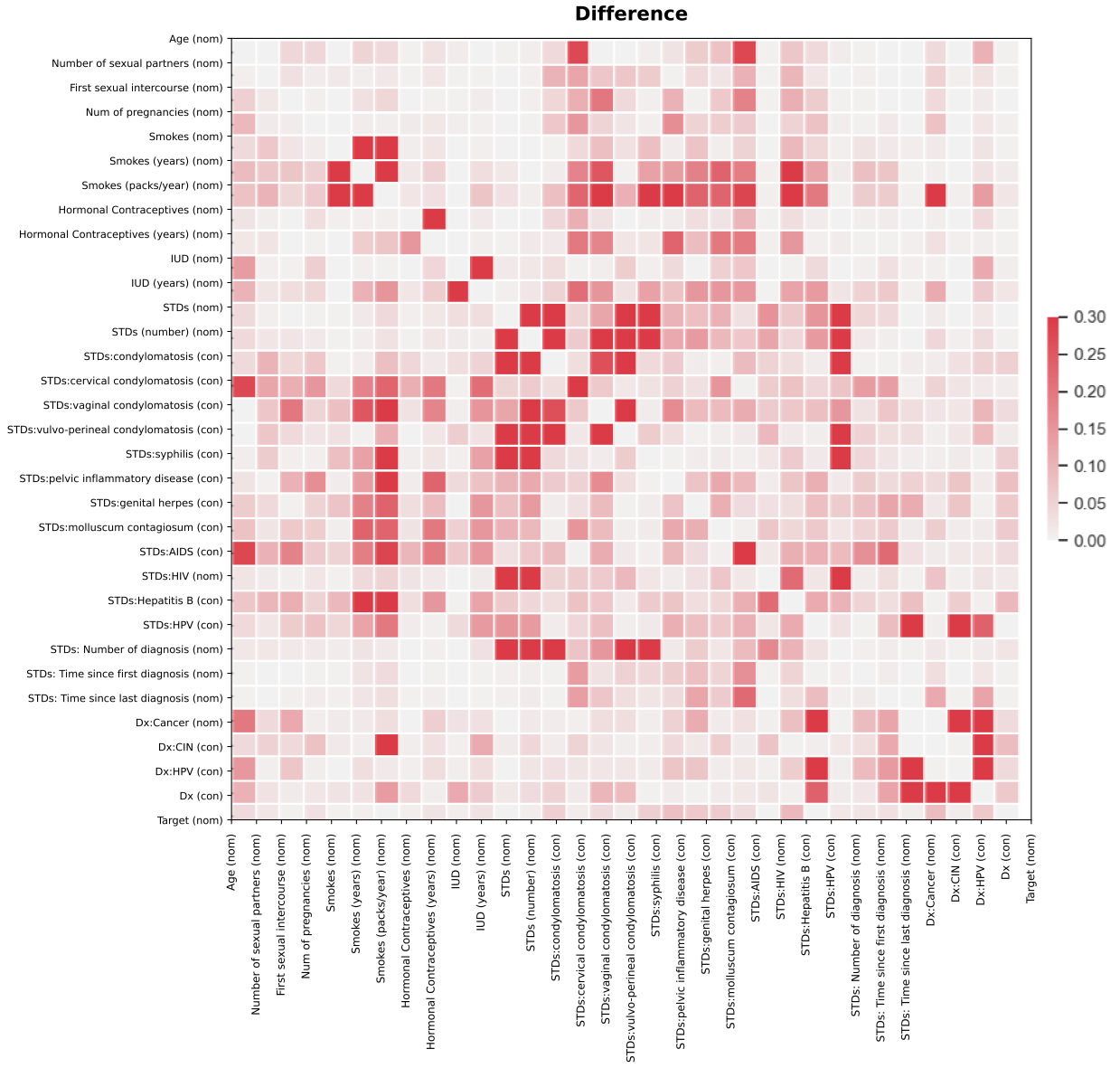

Figure 4.5: Comparison of Real and Fake Heatmap.

4.2.3 Forest Diffusion

The forest diffusion model represents an innovative approach to modeling tabular data, harnessing the strengths of decision tree architectures alongside diffusion processes. By integrating these techniques, the model effectively captures complex, non-linear relationships

inherent in structured datasets. This synthesis allows for improved predictive accuracy and robustness in various applications, from financial forecasting to healthcare analytics (Jolicoeur-Martineau et al., 2023).

4.2.4 Language Models are Realistic Tabular Data Generators

The Generation of Realistic Tabular data (GReaT) framework presents an innovative approach for synthesizing tabular data using language models (Borisov et al., 2023). Using the natural language processing capabilities of these models, GReaT efficiently generates realistic and coherent data entries that mimic the structure and correlations found in real-world datasets generated based on large language models based on the GReaT framework. The model is configured on DistilGPT-2 with a batch size of 32 for 16 epochs in a single run. This method offers a significant advancement in the field of data augmentation and synthetic data generation, providing a robust tool to improve model training and analysis without compromising data privacy. The mathematical formulation typically involves maximizing the likelihood of the generated sequences (rows) given the training data, which can be represented as:

$$\max \left(\sum_{i=1}^{N} \log P(\text{row}_i \mid \text{model}) \right) \tag{4.1}$$

where $P(\text{row}_i \mid \text{model})$ is the probability of observing the data in row i given the model.

4.3 Machine Learning Methods and Parameters

The dataset under investigation features a labeled target class, which requires the use of a supervised machine learning approach. The logistic regression classifier is a statistical model that uses a logistic function to model a binary dependent variable. It estimates the probability that a given input belongs to a particular category (Hosmer et al., 2013). Decision

tree classifier uses a tree-like model of decisions and their possible consequences (Quinlan, 1986). Random forest (RF) classifier operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes of the individual trees. This ensemble method is particularly noted for its high accuracy, robustness, and ease of use (Breiman, 2001). A gradient boost classifier combines several weak learning models to create a strong predictive model. Decision trees are typically used as weak learners in this approach. Gradient boosting is effective for a variety of prediction tasks (Friedman, 2001). A Gaussian naive Bayes classifier assumes that the continuous values associated with each feature are distributed according to a Gaussian distribution, which is particularly suitable for classification tasks in which the features are normally distributed (Murphy, 2006). k-nearest neighbors assumes that similar things exist in close proximity, it looks for the closest match of the test data in feature space. KNN is widely used for its simplicity and effectiveness in classification tasks (Altman, 1992). Support vector classifiers are known for their robustness and are particularly useful in scenarios where the number of dimensions exceeds the number of samples, suitable for many medical and biological classification tasks (Cortes and Vapnik, 1995). Multi-layer perceptrons are a class of feedforward artificial neural networks. MLP uses backpropagation for training the network, effective for complex non-linear modeling that captures intricate relationships in the data (Rumelhart et al., 1986).

4.4 Results

In the construction of models, various supervised learning techniques were employed to evaluate the effectiveness of data augmentation in clinical diagnostics. To ensure statistical reliability and prevent data leakage, a stratified 10-fold cross-validation was implemented. To maintain strict experimental integrity, the original dataset was first partitioned into 10 folds of the same size before the introduction of any synthetic samples. In each iteration,

the synthetic samples were concatenated exclusively with the training folds (90%), while the evaluation was performed solely on the original, unseen data (10%) stored in the testing fold.

This configuration ensures that the evaluation metrics reported represent the performance of the model on 100% authentic data, confirming that the synthetic generation process did not bias the final assessment. The performance metrics reported represent the mean and standard deviation across these ten iterations, providing a robust estimate of the generalizability of the model. All classifiers were implemented using the scikit-learn library (Buitinck et al., 2013), utilizing a standardized pipeline that included feature scaling to ensure compatibility across both linear and non-linear architectures.

Table 4.3 presents the evaluation metrics for the classifiers, highlighting the comparative performance between the baseline and the augmented datasets. For the baseline models, logistic regression achieved the highest recall (0.42 ± 0.15), while gradient boost, KNN, and MLP recorded the highest accuracy (0.86). The random forest baseline model achieved an accuracy of 0.84 ± 0.02, a precision of 0.27 ± 0.15, a recall of 0.16 ± 0.10, and an ROC-AUC of 0.64 ± 0.08. Following data augmentation, the random forest model recorded an accuracy of 0.89 ± 0.02, a precision of 0.61 ± 0.16, a recall of 0.46 ± 0.11, and an ROC-AUC of 0.80 ± 0.06. Logistic regression showed a change in ROC-AUC from 0.62 ± 0.10 to 0.66 ± 0.09, while its recall changed from 0.42 ± 0.15 to 0.47 ± 0.17. The decision tree classifier showed an increase in F1-score from 0.22 ± 0.15 to 0.45 ± 0.11.

4.5 Discussion

The results demonstrate that the inclusion of synthetic data provides a significant boost in the models' ability to navigate severe class imbalance. Although traditional models like logistic regression showed modest improvements in ROC-AUC, they remained constrained by their linear nature. In contrast, the ensemble and tree-based classifiers exhibited a trans-

Dataset	Classifier	Accuracy	Precision	Recall	F1-Score	ROC AUC
Original	Logistic Regression	0.73 ± 0.06	0.23 ± 0.10	0.42 ± 0.15	0.29 ± 0.11	0.62 ± 0.10
	Decision Tree	0.77 ± 0.05	0.19 ± 0.13	0.26 ± 0.18	0.22 ± 0.15	0.55 ± 0.10
	Random Forest	0.84 ± 0.02	0.27 ± 0.15	0.16 ± 0.10	0.20 ± 0.12	0.64 ± 0.08
	Gradient Boost	0.86 ± 0.02	0.32 ± 0.33	0.08 ± 0.06	0.12 ± 0.10	0.57 ± 0.09
	KNN	0.86 ± 0.02	0.30 ± 0.24	0.10 ± 0.09	0.15 ± 0.13	0.61 ± 0.07
	MLP	0.86 ± 0.03	0.42 ± 0.18	0.22 ± 0.12	0.28 ± 0.13	0.59 ± 0.11
Combined	Logistic Regression	0.71 ± 0.05	0.21 ± 0.08	0.47 ± 0.17	0.29 ± 0.11	0.66 ± 0.09
	Decision Tree	0.84 ± 0.04	0.41 ± 0.11	0.52 ± 0.13	0.45 ± 0.11	0.71 ± 0.07
	Random Forest	**0.89 ± 0.02**	**0.61 ± 0.16**	**0.46 ± 0.11**	**0.51 ± 0.08**	**0.80 ± 0.06**
	Gradient Boost	0.88 ± 0.02	0.64 ± 0.39	0.16 ± 0.11	0.25 ± 0.16	0.69 ± 0.08
	KNN	0.87 ± 0.03	0.49 ± 0.25	0.25 ± 0.13	0.32 ± 0.16	0.69 ± 0.08
	MLP	0.88 ± 0.03	0.62 ± 0.36	0.26 ± 0.16	0.35 ± 0.19	0.69 ± 0.08

Table 4.3: Classifier Performance: Original vs. Combined Datasets (10-Fold CV, Mean ± SD %)

formative response. The random forest model, in particular, emerged as the top performer, achieving a substantial increase in both recall (from 0.16 to 0.46) and ROC-AUC (from 0.64 to 0.80). This suggests that the synthetic data effectively modeled the minority class manifold, allowing the classifier to identify nearly triple the number of high-risk cases compared to the baseline without a significant loss in precision.

The analysis indicates that the proposed data augmentation strategy successfully addresses the limitations of sparse medical datasets. Although the original baseline models struggled with low recall-often missing more than 80% of risk cases the models trained on the combined dataset developed a more comprehensive understanding of characteristics of the minority class. Random forest emerged as the most reliable diagnostic tool, balancing a precision of 61.0% with a marked improvement in ROC-AUC. These results underscore the viability of using advanced generative techniques to enhance the predictive accuracy and clinical utility of cervical cancer screening models.a

4.6 Conclusions and Future Work

This study has highlighted the critical importance of cervical cancer classification while acknowledging the constraints posed by data availability and privacy concerns. Our research demonstrates that synthetic tabular data can effectively mitigate the challenges of severe class imbalance, which often leads to poor diagnostic sensitivity in baseline models. The results indicate that synthetic data do not necessarily replicate the limitations of the original dataset but actively improve the ability of the model to generalize, particularly in ensemble-based classifiers. Specifically, the marked improvement in recall and ROC-AUC suggest that the generative process successfully captured the underlying risk-factor correlations necessary to identify high risk patients.

Given the inherent limitations of a small, imbalanced clinical sample, the results should be interpreted as indicative of the augmentation's potential rather than definitive clinical proof. Validation on larger, multi-center datasets would be a necessary step to strengthen these conclusions.

In future work, we intend to refine the data preparation pipeline by implementing advanced dimensionality reduction and feature selection to isolate key risk factors before the generative phase. Furthermore, we propose investigating post-generation filtering techniques to remove low-fidelity synthetic outliers. Preliminary tests suggest that such refinements could further stabilize model performance across linear and non-linear architectures. Additionally, exploring the integration of multi-model synthetic ensembles and deeper neural architectures may provide further gains in predictive reliability.

4.7 Contributions of the Chapter

- **Data Scarcity Imbalance Paradox**

  - **Improve classifier performance under imbalance**: Synthetic data improves recall and F1 across multiple classifiers.
  - **Preserve multivariate relationships**: Synthetic data maintains clinical structure and improves several metrics.
  - **Optimize integration of synthetic + original data**: Synthetic dataset consistently outperforms original data.

- **Deployment Dilemma**
  - **Efficient models for low-resource settings**: Traditional ML models trained on synthetic data achieve strong performance with minimal computational cost.

Chapter 5

Tabular-to-Text Transformation and LLM-Based Classification

This chapter outlines the proposed methodology for developing the medical text classification model. It covers key steps including dataset description, preprocessing, model training, fine-tuning strategies, handling class imbalance, and addressing challenges faced during the process. The chapter concludes with an evaluation of the final model's performance. This study is published in IEEE International Conference on Prognostics and Health Management (ICPHM) (Gheibi and Hougen, 2025b).

5.1 Proposed Method and Dataset Description

For this study, we utilized the Cervical Cancer Risk Factors (CCRF) dataset from the UCI Machine Learning Repository (Fernandes et al., 2017). After the preprocessing phase described in Chapter 4, this dataset contains 30 features, a target value, and 753 patient records. The dataset provides both categorical and numerical attributes, which are transformed into

a textual format for processing by language models. Models are trained to predict whether a patient is likely to be recommended for further testing, with labels indicating "Yes" or "No." A label of "Yes" means that further testing will be recommended, while "No" means that no further testing is recommended.

5.2 Preprocessing and Data Transformation

Before training the model, the dataset undergoes a rigorous preprocessing phase. These preprocessing and data transformation steps are essential to ensure that the data is suitable for the classification model. By addressing data quality and consistency, we aim to improve the model's reliability and fairness, making it more applicable to real-world healthcare scenarios. The preprocessing phase includes the following steps:

- **Data Collection and Cleaning:** The preprocessing pipeline followed the same procedures described in Chapter 4, including handling missing values (e.g., mean or KNN imputation), and standardization of inconsistent entries. These steps ensured that the training data were clean, reliable, and aligned with the previously established methodology.

- **Data Splitting and Cross-Validation:** We performed k-fold cross-validation, in which the data was split into k subsets ($k = 10$) and the model was trained and validated across all fold combinations. This approach was particularly appropriate given the limited dataset size: a simple train-test split would have reserved a substantial portion of the data exclusively for testing, reducing the amount of data available for training, and increasing the variance of the performance estimate. In contrast, k-fold cross-validation allowed every sample to serve as both training and validation data across different folds, resulting in a more stable and reliable assessment of generalization

and reducing the risk that model performance depended on an unrepresentative split. This procedure ensured that the model generalized well to unseen data and did not overfit to any particular subset.

- **Tabular-to-Text Transformation:** Inspired by Wang et al. (2022), a work on converting structured data into text, we reformat tabular features into a structured textual representation that can be fed into a language model for classification.
- **Normalization of Text:** Feature values are converted into readable text where applicable and are articulated in natural language.

An example transformation is shown below:

> *"She is 30 years old. She has had 2.0 sexual partners. She has been pregnant 3.0 times. She smokes 3.5 packs per year. She has IUD for 9.0 years. She has used hormonal contraceptives for 3 years. She has 0.0 number of diagnosis. She has 0 history of STDs. The total number of STD diagnoses is 0.0. Time since the first diagnosis is 0.0 years. Time since the last diagnosis is 0.0 years. ...." "Label: No."*

5.3 Model Training and Fine-Tuning

This section describes the modeling strategies used to classify patients, focusing on both prompt-based inference with large language models and supervised fine-tuning of smaller, domain-specific models.

To classify patients as recommended for further testing or not, we applied both prompt-based inference using general-purpose language models and supervised fine-tuning of domain-specific small language models. The training was conducted in two phases:

1. **Zero-shot and Few-shot Learning:** General models Llama (Meta LLama Team, 2024), Mistral (Mistral AI, 2023), and Phi (Microsoft, 2024) were tested without additional training to establish baseline performance.

2. **Full Fine-Tuning:** We fine-tuned general purpose SLMs such as BioBERT (Lee et al., 2020), PubMedBERT (Gu et al., 2020b), ALBERT (Lan et al., 2019a), and DistilBERT (Sanh et al., 2019a) on the CCRF dataset, leveraging synthetic augmentation and class-weighted loss for improved cancer classification.

5.3.1 Fine-Tuning Strategy

The Fine-tuning of biomedical SLMs was performed using the Hugging Face trainer API (Hugging Face, 2024). Hyperparameters were selected through a small-scale grid search over learning rates $\{1 \times 10^{-5}, 5 \times 10^{-6}, 2 \times 10^{-6}\}$, batch sizes $\{8, 16\}$, and weight decay values $\{0, 0.01\}$, selecting the combination that maximized the validation F1-score. Each model was fine-tuned once using the selected hyperparameters. The final configuration is summarized below.

- **Learning Rate**: 2×10^{-6}
- **Epochs**: 10
- **Batch Size**: 16
- **Weight Decay**: 0.01
- **Optimizer**: Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$)
- **Model Selection**: Best validation F1-score

5.3.2 Handling Class Imbalance with Weighted Loss

The CCRF dataset exhibited a substantial class imbalance, with "No" cases comprising approximately 90% of the samples and "Yes" cases only 10%. This imbalance risked biasing the model towards always predicting the majority class, which would yield deceptively high accuracy while severely degrading recall for the minority "Yes" class. To mitigate this issue, we employed a weighted cross-entropy loss, where class weights were set inversely proportional to their frequencies in the training data. This weighting increased the penalty for misclassifying "Yes" cases, ensuring that the model learned to recognize minority-class patterns rather than defaulting to majority-class predictions.

5.4 Addressing Key Challenges

Initial experiments revealed suboptimal performance due to the usage of pronouns, implicit gender associations, and inconsistencies in feature representation. To improve input representation and consistency of the model, we implemented the following improvements:

- **Normalization of Features:**
  - Replaced raw numeric values with readable text for better interpretability (e.g., "three sexual partners" instead of "3.0").
  - Ensured uniform phrasing of categorical variables. Example: Instead of "1" or "0" for smoking status, we use "smokes" or "does not smoke."
  - Avoid using mean values to fill in missing data, as this can result in excessively long and non-intuitive numerical values (e.g., "First diagnosis is 27.0239"). Instead, we use more interpretable alternatives, such as the median or nearest whole number, to maintain readability and consistency in the text prompts.

- **Age Grouping:** Patient ages are categorized into text-based descriptors (e.g., 30.0 years old and younger as "young," older than 30.0 but younger than 50.0 as "middle-aged," and 50.0 and older as "older") to improve interpretability.

- **Ensuring Consistency in Text Prompts:** The text descriptions are refined to ensure uniformity across samples, mitigating potential implicit biases due to gender associations or pronoun usage.

- **Feature Grouping:** We categorized related features into meaningful groups to improve natural language understanding and model interpretability. For example, attributes such as the number of sexual partners, age at first intercourse, number of pregnancies were grouped under *Sexual history*, while all attributes related to sexually transmitted diseases (STDs) were grouped under *History of STDs*. This structured organization of information allowed for more coherent and contextually relevant text prompts, reducing inconsistencies, and improving the model's ability to capture relationships between risk factors.

- **Feature Selection:** To streamline the text prompts and make them more manageable for the model, we employed random forest for feature selection. Although random forests randomly sample subsets of features when constructing each tree, they produce aggregated importance scores based on how much each feature reduces impurity across the ensemble. These scores allow the model to identify which attributes contribute most strongly to the prediction task. This technique identifies the most significant features that contribute to the prediction task (Breiman, 2001). We ranked all features according to their random forest importance scores and performed a sensitivity analysis using the top 10, 15, 20, 25, and 30 features. The configuration with 20 features achieved the best validation performance. By removing low-importance features, the

resulting text prompts become shorter and more focused, improving both the training efficiency and perplexity in the language-modeling stage (Marvin et al., 2023; Tay et al., 2020).

- **Adjusting Decision Thresholds**: Given the imbalanced dataset, a default classification threshold of 0.5 led to poor recall for minority class cases. Instead, we dynamically adjusted the decision threshold to improve sensitivity to cancer cases.
- **Evaluation with precision-recall Metrics:** Instead of relying solely on accuracy, we monitored precision, recall, and F1-score to ensure effective classification.

An example transformation after modification is shown below:

"Patient Information:
*- **Age:** forty-four years (middle-aged adult).*
*- **Sexual history:** two sexual partners, first intercourse at age twenty-five. Pregnancy history: two pregnancies.*
*- **Smoking status:** smokes for (zero years, zero packs/year).*
*- **Contraceptive use:** using hormonal contraceptives for five years, IUD usage: using IUD for zero years.*
*- **History of STDs:** yes, Total STDs diagnosed: zero, Number of diagnoses: zero. Last STD diagnosis: zero years ago, First STD diagnosis: zero years ago...."*
***"Label:** No "*

5.5 Prompt Design

This section presents the design of prompts used for prompt-based classification of medical text. It introduces three prompting strategies—zero-shot, few-shot, and chain-of-thought

(CoT)—and explains their roles in evaluating model reasoning, adaptability, and performance in medical classification tasks.

In this study, we used prompt-based learning to classify medical texts into two categories: "Yes" and "No." The prompt design is critical to guide the model in correctly interpreting and classifying medical information. We employed three approaches: zero-shot, few-shot, and CoT.

5.5.1 Zero-Shot Prompt

In the *zero-shot* approach, the model is provided with a task description and a single medical text to classify, without any prior examples. The prompt explicitly instructs the model to classify the text into one of two categories: "Yes" or "No." The following is the structure of the zero-shot prompt:

> "You are an AI assistant trained for medical text classification. Your task is to classify the following medical text into one of two categories: "Yes" or "No".
>
> Rules:
>
> - Provide only one word as output: "Yes" or "No".
> - Do not provide explanations, reasoning, or additional text.
> - Your response must be exactly "Yes" or "No".
>
> Patient text: `{text}`"

The goal of this prompt is to obtain a direct binary classification from the model based solely on the medical context provided in the input `{text}`. Although prompts that request explanations or chain-of-thought reasoning are common in other NLP tasks, they introduce unnecessary variability for a strict Yes/No classification problem and can lead to inconsistent

or overly verbose output. Because the objective here is not to elicit interpretive reasoning but to evaluate the model's ability to make a categorical decision in a zero-shot setting, the prompt explicitly restricts the output to a single word ("Yes" or "No"). This constraint ensures consistency across predictions, reduces the risk of hallucinated explanations, and aligns the model's behavior with the requirements of a binary clinical classification task.

5.5.2 Few-Shot Prompt

In the *few-shot* approach, the model is given several examples of classified medical texts before being asked to classify a new text. This method helps the model learn from these examples and apply similar reasoning to new cases. The few-shot prompt includes a series of example medical texts along with their corresponding labels, followed by the new patient text that needs to be classified. An example of a few-shot prompt is shown below:

> "You are an AI assistant trained for medical text classification. Your task is to classify the following medical text into one of two categories: "Yes" or "No".
>
> Rules:
>
> - Provide only one word as output: "Yes" or "No".
> - Do not provide explanations, reasoning, or additional text.
> - Your response must be exactly "Yes" or "No".
>
> Below are some examples of correct classifications:
>
> Example 1: Text: `(Example 1 Text)` Label: `(Example 1 Label)`
>
> Example 2: Text: `(Example 2 Text)` Label: `(Example 2 Label)`
>
> Now, classify the following patient text as either "Yes" or "No". Do not provide explanations. Only return "Yes" or "No". Here is the text: `{text}`"

This approach provides the model with a few examples, which helps it better understand the classification task and improves its performance on unseen text. By presenting examples with labels, we can guide the model to follow the correct reasoning process when classifying new medical texts.

The design of the prompts is crucial for the success of prompt-based learning models, especially in medical text classification, where the texts contain complex terminology and domain-specific knowledge. By carefully crafting the prompts to be clear and concise, we ensure that the model can focus on the relevant information and generate accurate classifications.

5.5.3 Chain-of-Thought (CoT) Prompt

Encoder only SLMs such as BioBERT, PubMedBERT, ALBERT, and DistilBERT are discriminative classifiers and therefore cannot generate step-by-step reasoning or intermediate explanations. These models operate by mapping an input sequence to a fixed length representation and predicting a label, without any generative decoding capability. In contrast, *chain-of-thought* (CoT) prompting explicitly asks models to articulate intermediate steps before producing a final answer. Previous work by Wei et al. (2022) demonstrated that CoT prompting substantially improves performance in complex reasoning tasks by enabling models to decompose problems into interpretable steps. Motivated by these findings, and to complement the encoder-based experiments, we additionally evaluated CoT prompting using the CoT prompt applied in our experiments:

> "You are an AI assistant trained for medical text classification. Your task is to classify the following medical text. Let us think step by step about the medical risk factors in this case. Explain your reasoning briefly. After reasoning, provide your final answer on a new line in the format: Final Answer: Yes or Final Answer:

No.

Below are some examples of correct classifications:

Example 1:

Text: `(Example 1 Text)` Label: `(Example 1 Label)`

Example 2: Text: `(Example 2 Text)` Label: `(Example 2 Label)` ... ”

Example of the model outputs:

Some risk (smoking + HIV), but diagnosis negative Final Answer: No High risk (many partners, early intercourse, smoking, HIV) Final Answer: Yes Some risk (smoking) Final Answer: No Very low risk, but Dx code = 1 Final Answer: Yes

5.6 Experimental Setup

In this section, we describe the experimental setup used to evaluate the performance of both general-purpose large language models and fine-tuned small language models.

5.6.1 Zero-Shot, Few-Shot, and Chain-of-Thought

To investigate the behavior of generative Large Language Models, we evaluated their performance across three prompting strategies: zero-shot, few-shot, and chain-of-thought. In the few-shot configuration, each model was provided with three labeled examples within the prompt before classifying the target input. We evaluated Llama-3.2-3B, Mistral-7B, and Phi-4 under a 10-fold cross-validation framework. Although these models were not fine-tuned, they performed inference on each validation fold independently. For the few-shot and CoT configurations, all examples included in the prompts were drawn exclusively from the corresponding training folds to prevent data leakage.

5.6.2 Fine-Tuned SLMs

Fine-tuned SLMs (ALBERT, BioBERT, DistilBERT, and PubMedBERT) were evaluated using stratified 10-fold cross-validation. In each iteration, 9 folds (90%) were used for training, and 1 fold (10%) was used for testing. This process was repeated ten times so that each fold served as the test set exactly once. To ensure fair evaluation and prevent data leakage, synthetic samples were included only within the training folds. The metrics reported represent the mean and standard deviation across all folds, providing a statistically reliable estimate of the performance of the model.

5.7 Experiments and Results

This section presents the classification performance of general-purpose LLMs and fine-tuned SLMs on the CCRF dataset. Performance is evaluated using accuracy, precision, recall, specificity, and F1-score in a 10-fold cross-validation framework.

Model	**Accuracy**	**Precision**	**recall**	**Specificity**	**F1-Score**
Zero-Shot (ZS)					
Llama-3.2-3B	*83.99 ± 0.9*	*24.21 ± 5.3*	14.94 ± 3.6	*93.53 ± 1.9*	18.44 ± 4.2
Mistral-7B	78.21 ± 1.4	18.60 ± 2.2	*28.40 ± 4.2*	84.44 ± 1.6	*22.36 ± 3.4*
Phi-4	72.81 ± 2.2	14.71 ± 2.1	22.10 ± 3.4	80.52 ± 2.5	17.68 ± 2.8
Few-Shot (FS)					
Llama-3.2-3B	85.48 ± 1.3	25.29 ± 3.3	18.28 ± 2.9	93.53 ± 1.4	21.11 ± 3.8
Mistral-7B	86.99 ± 0.5	*35.78 ± 4.5*	*26.69 ± 4.1*	94.24 ± 0.6	*30.37 ± 4.9*
Phi-4	*87.54 ± 0.2*	18.30 ± 2.1	9.44 ± 1.8	*95.63 ± 0.2*	12.45 ± 2.2
Chain-of-Thought					
Llama-3.2-3B	87.35 ± 0.9	38.11 ± 4.5	31.00 ± 5.3	94.03 ± 1.0	34.20 ± 5.6
Mistral-7B	88.75 ± 1.9	44.22 ± 5.4	39.66 ± 6.1	94.32 ± 2.0	41.67 ± 6.5
Phi-4	**89.09 ± 1.7**	**48.11 ± 6.5**	**42.74 ± 6.9**	**94.56 ± 1.8**	**45.10 ± 5.9**

Table 5.1: Comparison of Zero-Shot (ZS), Few-Shot (FS), and Chain-of-Thought LLM Performance (10-Fold CV, Mean ± SD %). Best Results in Each Prompt Category is in *italics* and Best Results Overall is in **bold**.

Table 5.1 presents the performance of general-purpose LLMs under zero-shot (ZS), few-shot (FS), and chain-of-thought (CoT) prompting strategies. In the zero-shot setting, Llama-3.2-3B achieved an F1-score of 18.44%, Mistral-7B achieved 22.36%, and Phi-4 achieved 17.68%. Recall values remained below 30% across all models. In the few-shot setting, Mistral-7B achieved the highest F1-score of 30.37%, followed by Llama-3.2-3B at 21.11% and Phi-4 at 12.45%. In the chain-of-thought configuration, all models exhibited improved performance. Phi-4 achieved the highest F1-score of 45.10%, followed by Mistral-7B at 41.67% and Llama-3.2-3B at 34.20%. Standard deviations across folds ranged up to $\pm 6.5\%$, with higher variability observed in F1-score.

Finally, we evaluated small language models (SLMs) specialized for text classification, including ALBERT, BioBERT, DistilBERT, and PubMedBERT, with 235M, 66M, 110M, and 110M parameters, respectively. These models were tested on the dataset using their pre-trained versions and further fine-tuned using our proposed method. To address the inherent challenges of the small, imbalanced dataset ($N = 753$), we integrate synthetic data generated through our proposed GAN-based and LLM-based augmentation methods described in Chapter 4 into the fine-tuning process. To ensure statistical rigor and prevent data leakage, the stratified 10-fold cross validation split was performed exclusively on the original clinical dataset. In each iteration, synthetic samples were utilized only within the nine training folds to help the models learn minority class features, while the tenth fold remained a purely original, held-out validation set. This methodology ensures that our reported metrics calculated as the mean (μ) and standard deviation (σ) across all ten folds reflect the models' ability to generalize to real-world clinical observations rather than synthetic artifacts. Table 5.2 presents the performance of fine-tuned small language models. BioBERT achieved the highest F1-score of 63.73%, followed by PubMedBERT with 46.57%, DistilBERT with 31.27%, and ALBERT with 10.23%. BioBERT also achieved the highest precision (67.86%)

and recall (62.22%) among all the evaluated models.

Model	Accuracy	Precision	Recall	Specificity	F1-Score
ALBERT	87.78 ± 1.1	36.67 ± 5.3	6.11 ± 2.5	**98.60 ± 0.4**	10.23 ± 3.9
BioBERT	**90.83 ± 2.7**	**67.86 ± 5.1**	**62.22 ± 5.7**	95.35 ± 1.5	**63.73 ± 4.5**
DistilBERT	88.30 ± 2.4	53.43 ± 6.1	23.78 ± 4.9	97.16 ± 1.1	31.27 ± 5.1
PubMedBERT	89.51 ± 2.2	67.54 ± 5.8	38.44 ± 5.2	97.21 ± 0.9	46.57 ± 4.9

Table 5.2: Classification Performance Comparison of Fine-Tuned SLMs (10-Fold CV, Mean ± SD %)

5.8 Ablation Studies

To evaluate the contribution of key components to the proposed methodology, we conducted ablation experiments by systematically removing individual elements of the pipeline. The ablation study was conducted on BioBERT, the best-performing model in our experiments, to isolate and evaluate the contribution of individual components under the most effective configuration. The goal of this analysis is to assess whether each component provides measurable performance improvements under the same evaluation conditions.

The following configurations were evaluated:

- **Full Model:** Includes structured tabular-to-text transformation and class-weighted loss.
- **Without Structured Transformation:** Uses raw or minimally processed text without feature grouping or normalization.
- **Without Class Balancing:** Removes weighted loss, using standard cross-entropy.
- **Without Both Components:** Removes both structured transformation and class balancing.

All configurations were evaluated using the same 10-fold cross-validation protocol described in the experimental setup. Performance was measured using accuracy, precision, recall, specificity, and F1-score to ensure consistency with the main experiments. The results in Table 5.3 show that removing structured tabular-to-text transformation leads to a decrease in F1-score, indicating a reduced ability to capture relationships among clinical features. Removing class balancing results in a noticeable drop in recall for the minority class, confirming its importance in addressing dataset imbalance. The combined removal of both components produces the lowest overall performance in all metrics. These findings suggest that both structured input representation and class balancing contribute to improved classification performance, particularly in identifying minority-class cases.

Configuration	**Accuracy**	**Precision**	**Recall**	**Specificity**	**F1-Score**
Full Model	**90.83 ± 2.7**	**67.86 ± 5.1**	**62.22 ± 5.7**	95.35 ± 1.5	**63.73 ± 4.5**
No Structure	88.94 ± 2.6	60.42 ± 5.4	51.33 ± 5.9	94.18 ± 1.7	55.12 ± 4.9
No Class Balance	89.67 ± 2.5	64.10 ± 6.0	42.85 ± 5.6	97.02 ± 1.0	51.22 ± 5.3
Without Both	87.91 ± 2.8	58.25 ± 6.3	36.40 ± 6.1	**97.81 ± 0.9**	44.74 ± 5.8

Table 5.3: Ablation Study Results for BioBERT (10-Fold CV, Mean ± SD %)

5.9 Discussion

This section provides interpretation of the experimental results and compares the performance of general-purpose large language models and fine-tuned small language models. The results indicate that general-purpose LLMs evaluated under zero-shot, few-shot, and chain-of-thought prompting strategies exhibit limited effectiveness in identifying minority-class instances in the CCRF dataset. Although CoT prompting improves performance relative to zero-shot and few-shot configurations, the overall F1-scores remain moderate, suggesting that prompt-based inference alone may not be sufficient for highly imbalanced medical

classification tasks.

A notable observation is that several LLMs exhibit a tendency to favor the majority class ("No"), resulting in reduced recall for the minority class ("Yes"). This behavior is consistent with known challenges in imbalanced classification, where models prioritize overall accuracy while failing to capture clinically significant positive cases. The relatively high variance observed in F1-scores across folds further indicates sensitivity to the limited number of positive samples in the dataset.

The use of prompt engineering, particularly chain-of-thought prompting, appears to improve the ability of LLMs to incorporate relationships among clinical features. However, these improvements remain constrained, suggesting that reasoning-based prompting does not fully compensate for the absence of domain-specific supervised training.

In contrast, fine-tuned SLMs demonstrate substantially improved performance, particularly in terms of recall and F1-score. Among the evaluated models, BioBERT achieves the highest performance in multiple metrics. This suggests that domain-specific pretraining, combined with supervised fine-tuning, allows the model to better capture clinically relevant patterns and distinguish minority-class instances more effectively.

A key finding of this study is the performance difference between prompt-based inference and supervised fine-tuning. Although prompt-based methods allow general-purpose models to perform classification without additional training, their performance remains limited in imbalanced medical settings. In comparison, fine-tuned SLMs benefit from exposure to labeled training data, enabling more stable and accurate decision boundaries.

The ablation study further supports the effectiveness of the proposed methodology. The results indicate that both structured tabular-to-text transformation and class balancing contribute to improved performance. Removing either component leads to a reduction in model effectiveness, particularly in terms of minority-class recall. These findings suggest that care-

ful representation of input and imbalance mitigation are critical to achieving reliable performance in medical classification tasks.

Additionally, the use of synthetic data augmentation within training folds appears to support improved generalization without introducing data leakage. By restricting synthetic samples to training splits only, the evaluation remains focused on real-world clinical data, enhancing the reliability of reported metrics.

Despite these improvements, several limitations should be noted. The relatively small dataset size constrains the ability to draw statistically strong conclusions about fine-grained performance differences between models. As a result, the findings should be interpreted as a proof of concept rather than definitive evidence of superiority.

Future work may extend this approach by evaluating larger and more diverse clinical datasets, exploring alternative augmentation strategies, and developing models capable of recommending specific diagnostic tests rather than binary classifications. Additionally, integrating interactive and multi-language capabilities may further enhance the applicability of such models in real-world healthcare settings.

5.10 Contributions of the Chapter

- **Bridging the Modality Gap**
  - A structured tabular-to-text transformation framework was developed to convert clinical tabular data into natural language inputs suitable for language models.
  - This representation enables language models to capture contextual relationships among risk factors that are not explicitly modeled in traditional tabular learning approaches.
- **Addressing Data Imbalance**

  - The combination of class-weighted loss and structured input representation improved the model's ability to identify minority-class ("Yes") cases.
  - Experimental results show improved recall and F1-score for fine-tuned SLMs compared to prompt-based LLM methods in an imbalanced setting.

- **Comparison of General-Purpose and Domain-Specific Models**
  - A systematic comparison between general-purpose LLMs and domain-specific SLMs was conducted under a unified 10-fold cross-validation framework.
  - The results indicate that the domain-specific fine-tuned models (e.g., BioBERT) achieve a higher recall and F1-score than prompt-based LLM approaches.

- **Evaluation of Prompting Strategies**
  - Zero-shot, few-shot, and chain-of-thought prompting strategies were evaluated in medical text classification tasks.
  - Chain-of-thought prompting improved performance relative to zero-shot and few-shot settings but remained limited compared to supervised fine-tuning.

- **Ablation-Based Validation**
  - Ablation studies demonstrate that structured text transformation and class balancing both contribute to improved classification performance.
  - Removing either component leads to reduced F1-score and minority-class recall.

Chapter 6

From Narrative to Knowledge: A Two-Phase Latent Semantic Framework for Clinical Record Categorization

This chapter presents the motivation and methodological framework developed to categorize a dataset containing more than $165,000$ medical records. Phase 1 investigates a density-based clustering approach using language model (LM) embeddings and highlights the challenges associated with the context-dependent and domain-specific nature of clinical narratives. To address these limitations, Phase 2 introduces a probabilistic multi-label framework based on latent Dirichlet allocation (LDA), which is used to project documents into a latent semantic topic space rather than enforcing rigid cluster assignments. This process produces a silver standard, the dictionary-based specialty indicators were used as weak supervision signals (details in Section 6.5.2), labeled dataset that substantially reduces label noise and class overlap, providing a more stable target for supervised learning compared to raw hospital-provided dictionaries.

6.1 Introduction and Motivation

The rapid growth of electronic health records has created a large volume of unstructured clinical text, including physician notes, diagnostic impressions, procedure descriptions, and specialty specific terminology (Seinen et al., 2025). Although these narratives contain significant clinical insight, they remain difficult to organize and analyze at scale. Healthcare systems frequently rely on inconsistent, incomplete, or excessively granular labeling schemes that do not capture the true clinical intent behind each record. As a result, a substantial portion of clinical documentation effectively become *dark data*—information that exists within institutional systems, but remains insufficiently structured for scalable retrieval and analysis (Obrik-Uloho et al., 2025). Categorizing medical records into clinically coherent groups is essential for several downstream tasks, including automated classification, specialty routing, population level analytics, quality assurance, and decision support (Wang et al., 2018a; Sheikhalishahi et al., 2020). However, existing categorization methods face important limitations. Traditional dimensionality reduction and topic modeling methods, such as principal component analysis (PCA), when applied to raw frequency counts, often struggle with the dense, overlapping vocabulary of clinical notes. However, reframing the objective from hard clustering toward probabilistic semantic representation enables these methods to capture latent clinical structure beyond explicit keyword overlap. We can repurpose these methods to uncover the latent clinical intent that surface-level keyword matching misses (George and Birla, 2018). This chapter addresses these gaps by developing a data driven framework for organizing unstructured clinical narratives. Through a combination of unsupervised exploration, taxonomy engineering, and supervised validation, we construct a consolidated higher-level specialty taxonomy referred to herein as the Super-specialty taxonomy (details in Section 6.5) and demonstrate its effectiveness using both traditional baselines

and a PubMedBERT multi label classifier. This approach transforms noisy and inconsistent hospital labels into a structured and clinically actionable representation of patient records.

6.2 Data Collection and Description

This section first describes the dataset, followed by the preprocessing pipeline and the two phases of the proposed methodology.

6.2.1 Data Source

We utilized a corpus of 167,034 patient case summaries extracted from PubMed Central (PMC) (Zhao et al., 2023), a publicly accessible archive of biomedical and life sciences literature. These summaries represent de-identified clinical cases published in medical journals, covering a wide range of medical specialties and conditions.

Effective clustering of clinical narratives requires careful preprocessing to reduce non-informative lexical noise while preserving medically meaningful signal. Raw clinical case summaries contain narrative scaffolding, administrative artifacts, demographic descriptors, workflow terminology, and reporting conventions that can dominate vector representations without contributing to specialty-specific semantic structure (Meystre et al., 2008; Hossain et al., 2023).

To address this challenge, an iterative corpus-driven vocabulary refinement strategy was used. This strategy is inspired by corpus-specific stop-word adaptation and iterative feature selection methods commonly used in information retrieval and biomedical text mining (Salton and Buckley, 1988; Ramos, 2003). The objective was to progressively remove non-informative terms while retaining domain-relevant medical vocabulary essential for downstream clustering.

6.2.2 Initial Corpus Characteristics

The dataset exhibits several typical characteristics of clinical narratives. The raw corpus of the `patient_summary` exhibited substantial length variability:

1. **Length Variability**: Document lengths range from 11 to 13,688 words, with an average of approximately 415 words. This variability introduces several challenges for vector-based modeling. Longer documents can dominate term frequency-inverse document frequency (TF-IDF) weighting schemes, embeddings may become biased toward narrative verbosity, computational resource requirements increase substantially, and density-based clustering methods can become unstable. A formal definition of TF-IDF is provided in Section 6.2.3.

2. **Medical Complexity**: Texts contain detailed clinical descriptions that include symptoms, examination findings, diagnostic test results, treatment plans, and results.

3. **Specialty Diversity**: Cases span multiple medical specialties, including cardiology, oncology, neurology, gastroenterology, and others.

4. **Temporal Information**: Many summaries include temporal sequences of symptoms, diagnostic evaluations, and treatment responses.

These characteristics motivated systematic preprocessing, which are discussed in the following sections.

6.2.3 Term Frequency–Inverse Document Frequency (TF-IDF)

To transform textual data into a numerical representation suitable for clustering, we employ the TF-IDF weighting scheme. Let $\mathcal{D} = \{d_1, d_2, \ldots, d_M\}$ denote a corpus of M documents,

and let t be a term in the vocabulary (Oren, 2002). The *term frequency* (TF) of term t in document d is defined as:

$$\mathrm{TF}(t, d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}, \tag{6.1}$$

where $f_{t,d}$ is the number of occurrences of term t in document d.

The *inverse document frequency* (IDF) is defined as:

$$\mathrm{IDF}(t) = \log\left(\frac{M}{|\{d \in \mathcal{D} : t \in d\}|}\right), \tag{6.2}$$

where the denominator represents the number of documents that contain the term t.

The TF-IDF weight is then given by:

$$\text{TF-IDF}(t, d) = \mathrm{TF}(t, d) \times \mathrm{IDF}(t). \tag{6.3}$$

This formulation assigns high weights to terms that are frequent within a document but rare across the corpus, making it particularly suitable for emphasizing discriminative lexical features in clinical tasks (Salton and Buckley, 1988). In the context of clinical text, TF-IDF helps mitigate the influence of common narrative terms while emphasizing medically relevant vocabulary. The effectiveness of TF-IDF is strongly influenced by the quality of the underlying vocabulary, which motivates the iterative refinement strategy described in Section 6.3.

6.3 Preprocessing Pipeline

The clinical text preprocessing pipeline implemented several transformations to standardize and clean the input documents. The pipeline aimed to suppress narrative scaffolding (e.g.,

"patient presented with," "history of") while retaining clinically meaningful descriptors. A deterministic cleaning function was implemented using compiled regular expressions and rule-based filtering. The following normalization steps were applied:

- **Case normalization:** All text was converted to lowercase.
- **Whitespace normalization:** Multiple spaces, tabs, and line breaks were reduced to single spaces.
- **Numeric filtering:** All numeric expressions were removed (e.g., "45-year-old," "0.125 mg") to avoid demographic leakage into feature space.
- **Measurement unit removal:** Domain-specific units (e.g., *mg, ml, kg, cm*) were removed.
- **Stop-word removal:** Standard English stop-words were augmented with a domain-specific stop-word list constructed iteratively.

Abbreviation Normalization Clinical abbreviations were expanded using a predefined mapping dictionary to improve semantic consistency. For example, "w/" $\rightarrow$ "with," "h/o" $\rightarrow$ "history of," and "s/p" $\rightarrow$ "status post". The complete abbreviation mapping is provided in Appendix A.

Token Exclusion Criteria A token was removed from the corpus if it satisfied at least one of the following conditions:

- Numeric filtering: tokens containing digits
- Unit filtering: membership in a predefined measurement unit list
- Stop-word filtering: membership in either standard or domain-specific stop-word lists

- High document frequency: appearing in more than 70% of documents
- Low document frequency: appearing in fewer than 0.5% of documents

Domain-Specific Stop-word Construction A domain-specific stop-word list was constructed to remove non-informative tokens common in clinical narratives, including demographic descriptors (e.g., *patient, male*), temporal markers (e.g., *day, month*), narrative verbs (e.g., *presented, showed*), and procedural terms (e.g., *treatment, procedure*). The use of domain-specific stop-word lists is important in specialized corpora, as generic stop-word lists may not capture high-frequency but semantically uninformative terms unique to a given domain. Removing such terms can reduce the dimensionality of the feature space and improve the performance of downstream text mining tasks. In clinical text, domain-specific preprocessing is particularly important due to the presence of specialized terminology, abbreviations, and variability in narrative structure (Makrehchi and Kamel, 2008; Alshanik et al., 2020).

6.3.1 Iterative Corpus-Driven Vocabulary Refinement

The stop-word vocabulary was refined iteratively using corpus-level frequency statistics and LLM-based inspection of non-informative words to ensure the removal of structurally frequent but clinically uninformative terms. The complete stop-word list is provided in Appendix A. The basis for classifying these tokens as non informative is that they serve grammatical, administrative, or descriptive functions rather than conveying clinical concepts. For example, scaffold verbs such as "show" or "perform" describe the act of documentation rather than the medical condition. Demographic terms such as "year" or "male" appear in almost every record and do not indicate a specific specialty. Time and unit markers such as "day," "month," "mg," or "ml" are routine measurement words that do not carry diagnostic or spe-

cialty specific information (Kim et al., 2012). Because these terms occur frequently in all specialties and do not reflect clinical intent, they were removed to reduce noise and allow the models to focus on medically meaningful vocabulary. Importantly, a specialty-bearing medical vocabulary was retained, including *tumor, liver, artery, bone, MRI, biopsy, abdominal, chest, serum* B.

We formalize preprocessing as an iterative corpus-driven vocabulary refinement Algorithm 1. The objective is to maximize the density of specialty signals while minimizing narrative and structural noise prior to statistical feature filtering and clustering.

Algorithm 1: Iterative Corpus-Driven Vocabulary Refinement

Data: Clinical corpus $\mathcal{D} = \{d_1, \ldots, d_M\}$; Initial stop-word set $\mathcal{S}_0$; Top-N threshold N; Convergence threshold τ

Result: Cleaned corpus $\mathcal{D}^*$

$\mathcal{S} \leftarrow \mathcal{S}_0$;

$converged \leftarrow false$;

while $converged = false$ **do**

  foreach $d_i \in \mathcal{D}$ **do**

    Apply preprocessing pipeline defined in Section 6.3;

    Tokenize, remove tokens in $\mathcal{S}$;

    Store cleaned tokens as d'_i;

  Construct vocabulary $\mathcal{V}$ from $\{d'_1, \ldots, d'_M\}$;

  Compute corpus frequency $f(w)$ for each $w \in \mathcal{V}$;

  Sort $\mathcal{V}$ in descending order of $f(w)$;

  $\mathcal{V}_{top} \leftarrow$ first N words of sorted vocabulary;

  $\mathcal{N} \leftarrow$ LLM-assisted semantic inspection of high-frequency tokens($\mathcal{V}_{top}$);

  Compute medical signal ratio:

$$r = \frac{|\mathcal{V}_{top} \setminus \mathcal{N}|}{N}$$

  if $r \geq \tau$ **then**

    $converged \leftarrow true$;

$\mathcal{D}^* \leftarrow \{d'_1, \ldots, d'_M\}$;

Save $\mathcal{D}^*$;

return $\mathcal{D}^*$;

As mentioned in Section 6.3.1, the identification of non-informative words ($\mathcal{N}$) was conducted using a large language model (GPT5) with a zero-shot prompting strategy that acts as a clinical linguistic annotator. For each iteration, the top N tokens (words) were formatted into a list and passed to the model. The model returned a binary classification for each token. Only tokens classified as scaffolding or noise were added to the expanded stop-word set $\mathcal{S}$. The system prompt utilized is as follows:

> *"You are a medical data scientist. Review the following list of clinical tokens. Classify each token as either* ***Signal*** *(specific physiological states, diagnoses, or treatments) or* ***Scaffolding*** *(narrative structure, common verbs, or reporting language). Return the results as a list of tokens to be excluded (Scaffolding)."*

6.3.2 Rationale for Top-N Frequency Inspection

Threshold values were selected on the basis of Zipfian coverage and empirical convergence behavior. Specifically, N was empirically selected so that the top-ranked vocabulary represented approximately 50% of the total corpus token frequency mass to focus on the most influential portion of the corpus. The convergence threshold was set at $\tau = 0.6$, indicating a sufficient dominance of the clinical signal in the high-frequency vocabulary and reflecting the diminishing returns of further refinement.

Natural language corpora follow Zipf's Law, where the frequency of word $f(w)$ is inversely proportional to rank r:

$$f(w) \propto \frac{1}{rank(w)}. \tag{6.4}$$

Consequently, a small subset of the vocabulary accounts for a disproportionately large share of the total corpus. Let cumulative coverage $C(N)$ be defined as the ratio of the

frequency of the top-N words to the total number of tokens in the vocabulary V:

$$C(N) = \frac{\sum_{i=1}^{N} f(w_i)}{\sum_{w \in \mathcal{V}} f(w)}. \tag{6.5}$$

Empirically, due to the Zipfian distribution, which is characterized by a strong concentration of probability mass in a small number of high-frequency items, the majority words occur rarely. Consequently, a relatively small subset of the vocabulary accounts for a disproportionate share of total corpus tokens. Thus, selecting N such that $C(N) \approx 0.5$ ensures that inspection focuses on the dominant lexical mass while maintaining tractability (Tullo and Hurford, 2003).

6.3.3 Transition to Statistical Filtering

Following iterative refinement, approximately 60% of the top-ranked tokens were identified as medically meaningful, indicating that a substantial portion of high-frequency terms had transitioned from general-purpose to domain-relevant usage. At this stage, further manual intervention yields diminishing returns and increases the risk of over-curation. Consequently, we transitioned to automated document frequency filtering to handle the remaining lexical distribution, using empirically motivated thresholds:

- **Upper Bound:** Terms appearing in $> 70\%$ of documents were removed as non-discriminative corpus-wide stop-words, consistent with common practices in information retrieval for removing corpus-wide stop-words.
- **Lower Bound:** Terms appearing in $< 0.5\%$ of documents were removed as sparse noise (e.g., rare misspellings or idiosyncratic entries).

This dual-threshold approach preserves domain-specific signals while reducing noise and

improving the stability of the resulting vector space representations for downstream clustering.

Corpus Characterization and Length Distribution The iterative cleaning process resulted in a substantial reduction in corpus volume, decreasing the average document length from approximately 415 words to 170 words per record (approximately 60% reduction). Length normalization was achieved through a combination of token-level filtering and mild document-length standardization.

Specifically, the removal of high-frequency scaffolding terms and document frequency filtering reduced disproportionate token accumulation in longer documents, thereby mitigating length-related bias in the feature space. Documents containing fewer than five tokens were excluded because ultra-short records provide insufficient semantic information for stable representation learning and clustering.

Overall, the corpus remained statistically stable after preprocessing. Although the maximum post-cleaning document length reached 7,082 tokens, the median length of 124 tokens and the relatively moderate standard deviation indicate that most records retained meaningful semantic content without substantial over-cleaning or sparsity. Together, these preprocessing steps reduced length variability while preserving clinically relevant information, ensuring that downstream representations were driven more by semantic content than by raw document size.

Table 6.1 summarizes the effect of the preprocessing pipeline on corpus structure. The substantial reduction in average and median document length indicates successful removal of narrative scaffolding and non-informative lexical content while preserving clinically meaningful terminology. The reduction in variance further suggests improved corpus uniformity, which is beneficial for stable vector representation and downstream clustering.

Metric	Before Preprocessing	After Preprocessing
Minimum Length	1 token	5 tokens
Median Length	302 tokens	124 tokens
Maximum Length	13,688 tokens	7,082 tokens
Mean Length	415 tokens	170 tokens
Standard Deviation	241 tokens	89 tokens

Table 6.1: Corpus Statistics Before and After Preprocessing

Medical Vocabulary Density Empirical inspection of the refined corpus indicates high medical signal density within the upper frequency ranks, particularly among the top 100 tokens, with consistently strong semantic relevance across the top 200 to 400 tokens. However, a distinct inflection point emerges beyond the 400-token threshold, where medical relevance begins to decline more noticeably. This indicates that narrative expansion and general linguistic structure increasingly dilute the domain-specific signal in lower-frequency tiers, justifying our focus on the top-N most frequent terms.

Before processing: *A* 58*-year-old woman with a* 2*-year history of polyarthropathy had a diagnosis of RA (Fig.). She was treated with oral corticosteroids (*15 *mg/QD), methotrexate (MTX)* 10 *mg weekly, and/or a nonsteroidal anti-inflammatory drug. Her steady situation lasted for* 16 *months. In the recent* 8 *months, she experienced severely impairing and dizzines s and anemia. The blood test revealed ....*

Extracted entities after processing: *[platelet, plt, erythrocyte, polyarthropathy, mcv, sedimentation, esr, crp, corticosteroids, methotrexate, impairing, dizzines, anemia, normocytic, laboratory, cell, wbc, hemoglobin, erythrocyte, corpuscular, volume, rheumatoid, factor, anti,....]*

The preprocessing pipeline transforms raw narrative text into a semantically denser, domain-adapted representation suitable for clustering. By combining iterative refinement in

Algorithm 1 with statistical filtering, the final corpus balances noise reduction and domain signal preservation, providing a stable and semantically enriched foundation for downstream representation learning and clustering experiments.

6.4 Phase 1

The initial strategy explored hybrid feature fusion between semantic embeddings and structured clinical representations. This section focuses on synthesizing high-dimensional semantic representations with structured clinical indicators to create an effective feature space of documents.

6.4.1 Semantic Feature Extraction with Language Models

We utilized PubMedBERT to generate semantic document embeddings. PubMedBERT, which is pretrained on biomedical literature, was chosen over general-domain BERT variants because of its superior performance on biomedical NLP benchmarks such as BioASQ and BLURB (Gu et al., 2021).

6.4.2 Structured Clinical Feature Extraction

A Multi-component approach was developed to extract structured clinical features from the document corpus, combining entity-based and specialty-based representations.

Medical Specialty Identification According to the American Board of Medical Specialties (ABMS) (American Board of Medical Specialties, 2026) 18 main medical specialties were defined with comprehensive keyword sets. Each specialty was associated with 10-15 clinically relevant keywords (detailed in Appendix B). For each document, a specialty vector was

calculated, where each element represented the normalized frequency of keywords associated with that specialty. This approach enabled the quantification of document specialization while accounting for multidisciplinary content.

Clinical Entity Extraction Named entity recognition (NER) was performed to identify clinically relevant concepts, including diagnoses, procedures, medications, and anatomical references. The extracted entities were transformed into a TF-IDF representation with 500 maximum features, applying minimum document frequency filtering (5 to 10 occurrences) to control feature sparsity and computational complexity and to eliminate rare terms while retaining clinically significant vocabulary. We implemented a hybrid entity extraction approach that combined:

- **Dictionary-based extraction** using medical terminologies
- **Pattern-based extraction** using regular expressions for clinical phrases
- **Statistical extraction** using TF-IDF for domain-specific term identification

Feature Integration The structured feature vector for each document was constructed by concatenating the specialty vector with the entity-based TF-IDF representation. This hybrid representation captured both document-level specialization patterns and fine-grained clinical concept distributions.

Hybrid Feature Fusion To integrate semantic language-model embeddings with structured clinical features, a two-stage alignment and fusion process was implemented. Given the mismatch between the language model embeddings (768 dimensions) and the structured feature set (518 dimensions), PCA was used to reduce embedding dimensionality while approximately aligning the representation scale with the structured feature space prior to

fusion.

Feature Normalization Both feature sets underwent z-score normalization using the StandardScaler implementation from scikit-learn. This transformation centered each feature dimension on zero mean and unit variance, eliminating scale disparities between embeddings (typically distributed in $[-1, +1]$) and structured features (ranging across multiple orders of magnitude). Normalization was applied independently to each feature set to preserve representation balance and prevent scale dominance during fusion.

Weighted Fusion A convex combination approach fused the normalized feature representations:

$$F_{\text{hybrid}} = (1 - \alpha) \cdot F_{\text{BERT}} + \alpha \cdot F_{\text{structured}} \tag{6.6}$$

where $\alpha \in [0, 1]$ controlled the relative contribution of structured clinical features. Hyperparameter tuning via grid search identified $\alpha = 0.3$ as a strong-performing configuration, balancing semantic information (70% weight) with clinical specificity (30% weight).

6.4.3 Clustering Methodology

In clinical applications, cluster quality must be assessed through both quantitative metrics and qualitative clinical relevance. The absence of ground truth labels necessitates careful selection of validation approaches that capture different aspects of cluster structure.

Evaluation Metrics for Clustering Five complementary clustering evaluation metrics were used, combining both internal validation criteria and external agreement measures based on reference specialty labels.

- **Silhouette Score** measures cluster cohesion and separation by comparing intra-cluster

distances to inter-cluster distances. The values range from -1 (poor clustering) to $+1$ (excellent clustering), with higher scores indicating better-defined clusters (Rousseeuw, 1987).

- **Davies-Bouldin Index** quantifies the average similarity between each cluster and its most similar counterpart, where lower values indicate better separation between clusters. This metric emphasizes compact, well-separated cluster formations (Chicco et al., 2025).

- **Calinski-Harabasz Index** (also known as the Variance ratio criterion) computes the ratio of dispersion between-clusters to within-cluster dispersion. Higher values correspond to denser, more separated clusters, with the metric being particularly sensitive to cluster size variations (Chicco et al., 2025).

- **Normalized Mutual Information (NMI)** measures the agreement between the predicted clusters and reference labels derived from clinical specialty groupings, normalized to scale the results between 0 and 1. We utilize NMI to assess how much information is shared between our clinical specialties and the resulting clusters, as it is invariant to the absolute values of the class labels (Vinh et al., 2010).

- **Adjusted Rand Index (ARI)** evaluates the similarity between two assignments by considering all pairs of samples and counting pairs that are assigned in the same or different clusters in the predicted and reference labels derived from clinical specialty groupings. The adjusted nature of the score accounts for the chance grouping of elements, making it a more rigorous measure of clustering accuracy in imbalanced clinical datasets (Hubert and Arabie, 1985).

These metrics collectively provided a multidimensional assessment of cluster quality, addressing different failure modes: silhouette score for general cluster definition, Davies-Bouldin

for separation quality, and Calinski-Harabasz for density and variance characteristics.

Algorithm Selection We selected k-means clustering as our primary clustering algorithm due to its computational efficiency, scalability to large datasets, and interpretability. Although there are more sophisticated clustering algorithms, k-means provides an effective baseline and has been widely used in clinical text clustering studies.

We used the optimal cluster count (k=18) on $10,000$ samples determined by silhouette analysis, normalized mutual information, and adjusted rand index. Table 6.2 presents the quantitative comparison between raw text, BERT embeddings, and our hybrid approach. Notably, the *Raw Data* achieved the highest silhouette score, while the *Hybrid* approach, which incorporates more medical structure, saw a decrease in traditional metrics.

	Method		
Metric	**Raw data**	**Embeddings**	**Hybrid**
Silhouette Score ↑	**0.391 ± 0.021**	0.351 ± 0.021	0.328 ± 0.021
Davies-Bouldin Index ↓	**0.998 ± 0.002**	1.04 ± 0.002	1.13 ± 0.002
Calinski-Harabasz Index ↑	**4555.3**	2751.6	2610.4
NMI Score ↑	**0.30 ± 0.02**	0.23 ± 0.02	0.25 ± 0.02
ARI Score ↑	**0.22 ± 0.02**	0.13 ± 0.02	0.16 ± 0.02

Table 6.2: Quantitative Clustering Performance Comparison (k=18 clusters, ↑ Higher is Better, ↓ Lower is Better)

Although silhouette and Davies-Bouldin indices are standard for spherical cluster validation, they proved sub-optimal for the PMC clinical corpus. These metrics tend to penalize the presence of natural clinical bridges between related specialties. For example, strong semantic similarity between related domains such as *Valvular Cardiology* and *Coronary Intervention* creates continuous semantic overlap rather than clearly separable Euclidean regions. This inverse relationship between metric quality and medical significance reveals a significant challenge in clinical NLP. High scores in raw data reflect a clustering of narrative

scaffolding and demographic templates (e.g., "X-year-old male presented with...") rather than clinical pathology. As the model shifts toward semantic embeddings, these template clusters dissolve into a continuous manifold of overlapping medical domains, naturally lowering Euclidean separation scores. When projected into low-dimensional visualization space, this produces a densely entangled projection structure (informally referred to as a "hairball" effect) in UMAP projections shown in Figure 6.1, where valid multidisciplinary overlaps are incorrectly interpreted as clustering noise. This limitation necessitated a move away from distance-based clustering toward a latent projection framework.

The hairball effect observed in the initial UMAP projections suggests that valid clinical overlaps such as the relationship between *Radiology* and *Oncology* are interpreted by standard metrics as poor separation. Consequently, the results in Table 6.2 should therefore be interpreted as evidence of intrinsic semantic overlap rather than poor model construction. This limitation necessitated a move toward a *latent semantic projection* (Phase 2, Section 6.5), which evaluates success based on label density rather than spatial isolation.

To resolve the Euclidean limitations of initial UMAP projections, we applied hierarchical cluster analysis (HCA) to the keywords refined via the iterative process in Algorithm 1. To maintain computational efficiency over a dataset of $N = 167,342$, we implemented a centroid-based abstraction. We first utilized mini-batch k-means to compress the high-dimensional TF-IDF space into ($k = 20,000$) representative semantic centroids to balance semantic granularity and computational tractability, capturing fine-grained sub-specialty variation present in the data that lower-resolution models fail to preserve. The HCA was then performed on these centroids. The cluster sizes indicated in parentheses in Figure 6.2 refer to the number of semantic centroids within each branch; for example, the largest leaf nodes represent high-density semantic clusters aggregating over $3,500$ original clinical records. As illustrated in Figure 6.2, the resulting dendrogram reveals an interpretable

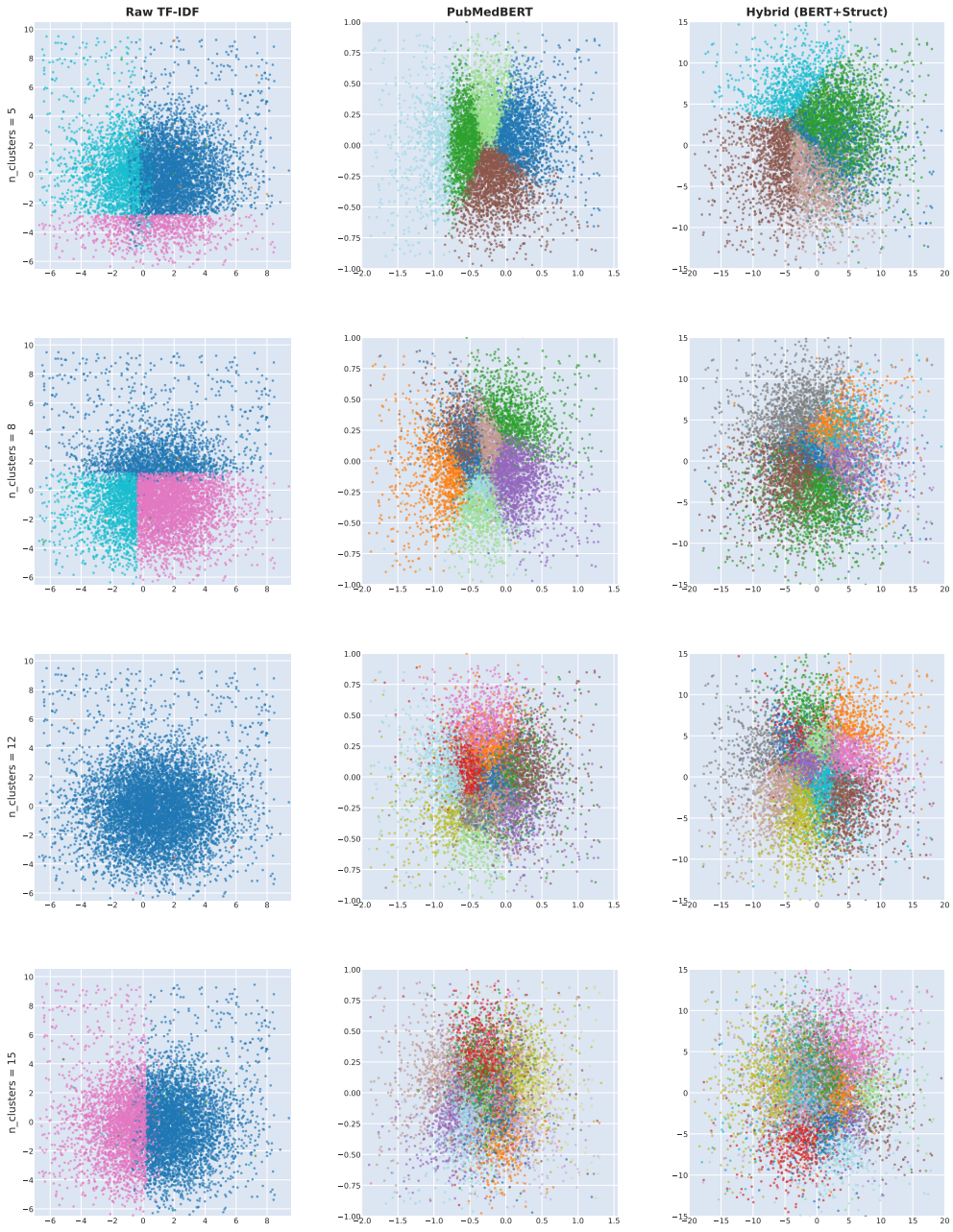

Figure 6.1: UMAP Visualization Illustrating Dense Semantic Overlap Among Clinical Embeddings

semantic structure previously obscured by clinical noise. However, the HCA results also provide empirical evidence of persistent clinical entanglement. The structural rigidity of HCA, which forces hard membership for inherently overlapping specialties, confirms that clinical specialties do not exist as isolated clusters but as nested semantic relationships. This limitation concludes Phase 1, indicating that rigid Euclidean or hierarchical approaches are insufficient to fully resolve the observed metric limitations.

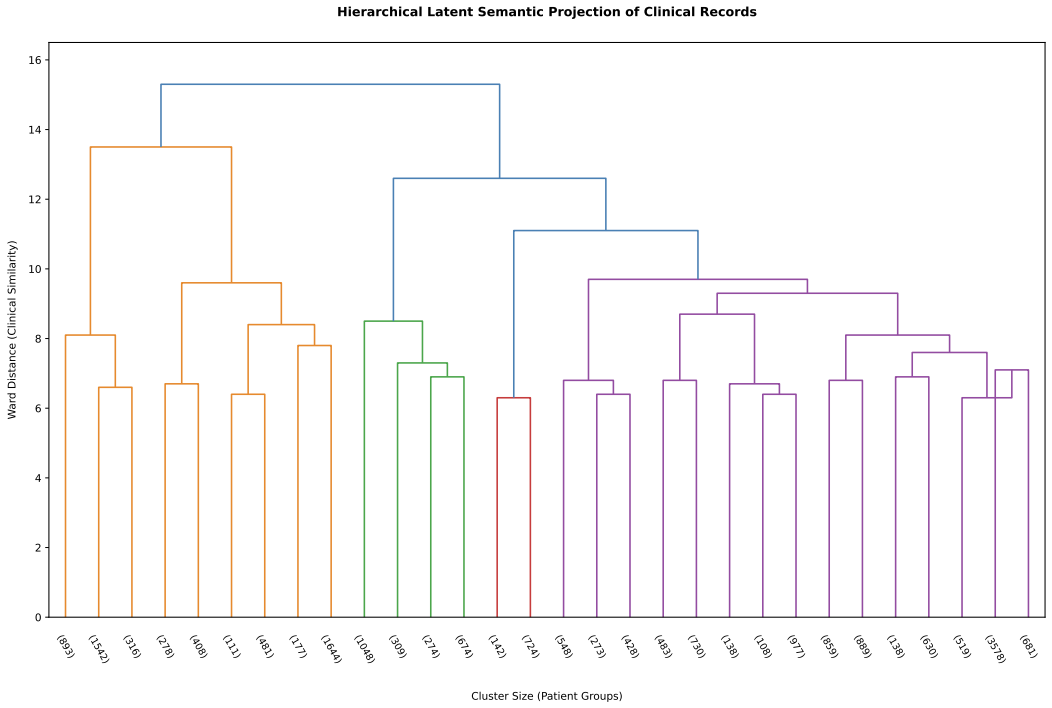

Figure 6.2: Hierarchical Cluster Analysis (HCA) Applied to the Refined Keyword Representations

6.5 Phase 2

As shown in Section 6.4, traditional k-means clustering ($k = 18$) and HCA yielded suboptimal separation between medical domains. These limitations revealed a metric paradox in which clinically meaningful overlap between specialties was penalized by distance-based clustering

metrics, motivated a transition to a probabilistic framework. This overlap is illustrated later in Figure 6.3. We propose a latent semantic projection approach using latent Dirichlet allocation to map medical records into probabilistic Super-specialty representations, allowing records to exist across multiple clinical intersections simultaneously.

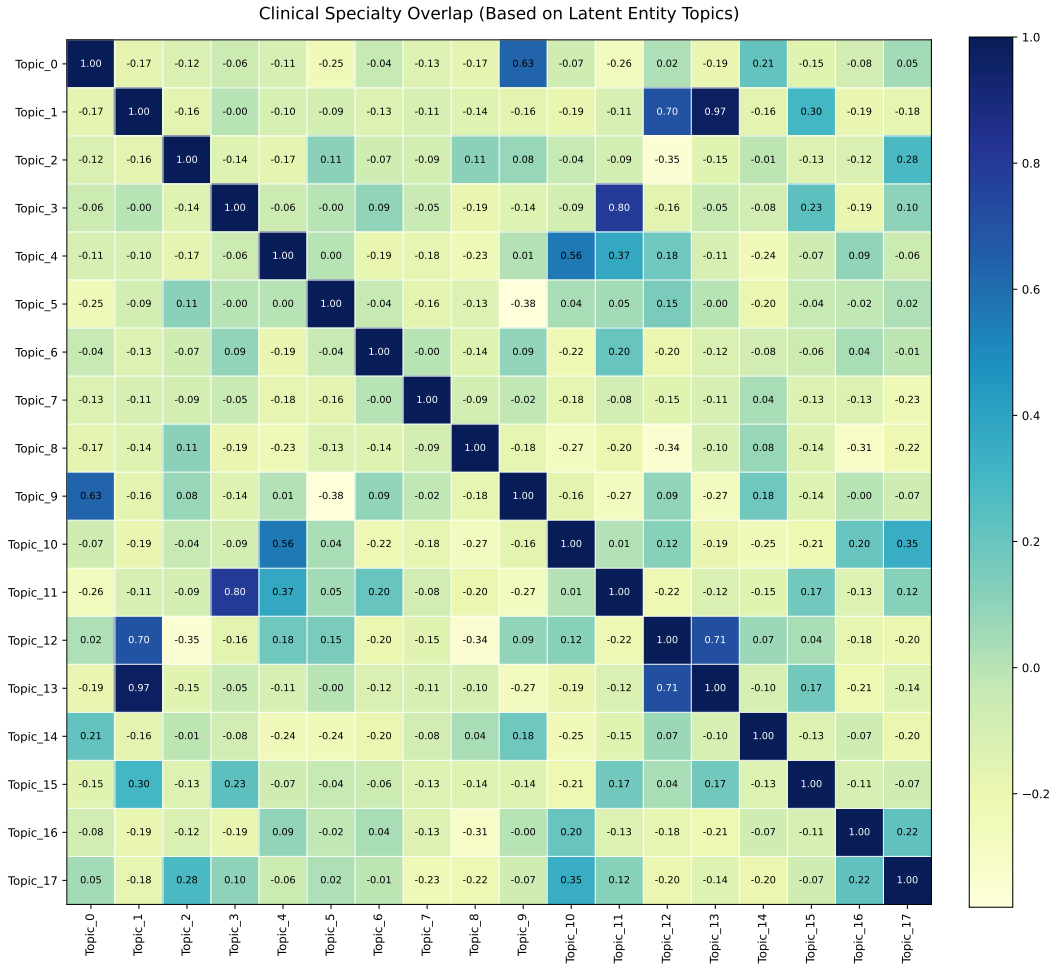

Figure 6.3: Visualization of Semantic Overlap Among Clinical Specialties in the Latent Topic Space Generated by LDA

6.5.1 Latent Semantic Projection (LSP)

To address the limitations of Euclidean distance-based clustering identified in Phase 1 (Section 6.4) and to better separate clinically meaningful structure from narrative scaffolding, we propose a latent semantic projection framework based on latent Dirichlet allocation to the medical entities extracted from the corpus. Unlike the distance-based representations used in Phase 1, LDA explicitly models each document as a distribution over K latent topics, where $K = 18$ was selected to approximately align with the higher-level specialty structure derived from the initial specialty taxonomy.

To reduce direct dependence on existing label frequencies and encourage topic discovery from latent corpus structure, the model was configured with a symmetric Dirichlet prior ($\alpha = \beta = 1/K$) and trained using batch variational inference. This prior encourages sparse but flexible topic mixtures, allowing records to maintain probabilistic membership across multiple specialties. This process transformed the high-dimensional entity space into a dense 18-dimensional *latent clinical space*. This projection captures the multi-label nature of clinical data, allowing a single record to exist simultaneously in multiple domains (e.g., 70% Cardiovascular, 30% Radiology) while maintaining stable probabilistic topic assignments.

6.5.2 Specialty Mapping and silver standard Projection

In clinical natural language processing, a *silver standard* refers to an evaluation target or proxy corpus generated via automated annotation pipelines, multi-system consensus algorithms, or weak supervision paradigms rather than exhaustive, manual adjudication by domain-expert clinicians (Rebholz-Schuhmann et al., 2010). To transform the latent topics into interpretable medical specialties, we established a mapping matrix $\mathbf{M}$, derived from the alignment between latent topics and dictionary-based specialty indicators. As described in

Phase 1 (Section 6.4), the dictionary-based specialty indicators were used as weak supervision signals to calibrate topic interpretation. For each record d, the topic distribution θ_d was projected onto the specialty space to generate a probability vector $\mathbf{S}_d$:

$$\mathbf{S}_d = \theta_d \cdot \mathbf{M}^\top. \tag{6.7}$$

where $\mathbf{M} \in \mathbb{R}^{K \times S}$ is the topic-to-specialty mapping matrix, $K = 18$ denotes the number of latent LDA topics, and S represents the number of consolidated *super-specialty domains*.

This projection enables a multi-label representation, where a single patient record can possess high confidence scores across multiple specialties (e.g., 0.7 Cardiovascular and 0.3 Radiology), mitigating densely overlapping semantic regions observed in Phase 1 (Section 6.4) by explicitly modeling overlapping specialties. The resulting topic structure enabled the consolidation of noisy categories into distinct super-specialty domains. This consolidation reduced linguistic redundancy between semantically related specialties (e.g., Surgery vs. Oncology) and providing a more coherent silver standard for downstream modeling.

Analysis of the topic correlation matrix (see Figure 6.4 in the Results) revealed substantial clinical bridges. To reduce linguistic redundancy and class imbalance, we performed a data-driven consolidation of the primary medical specialties into super-specialty domains.

- **Example:** Topic 1 (Valvular Heart Disease) and Topic 13 (Coronary Intervention) exhibited strong topic correlation (e.g., Pearson $r \approx 0.97$ computed over document-level topic probability vectors), suggesting that they can be consolidated into a single *Cardiovascular* domain.

This consolidation produced a more interpretable and semantically stable silver standard, which served as the final target representation for downstream supervised learning tasks.

6.6 Experimental Setup

The preprocessing pipeline was implemented in Python 3.9 using libraries such as *NLTK, spaCy, SciSpaCy* with the *en-core-sci-lg* and *en-ner-bc5cdr-md* models, and the regular expressions module for pattern matching and string manipulation operations. Each document was subjected to this preprocessing sequence before feature extraction to ensure consistent input quality across all subsequent pipeline stages. This stage reduced superficial lexical variability while preserving medical terminology. We used PubMedBERT, a domain-specific model pretrained on biomedical literature, to generate semantic embeddings of clinical documents. PubMedBERT was selected over general-domain BERT models due to its strong performance in biomedical NLP tasks, including benchmark evaluations such as BLURB (Gu et al., 2021).

The embedding generation process comprised three sequential steps. First, the model was initialized using the "BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext" configuration with mean pooling for sentence-level embeddings. Second, documents were processed in batches of 32 for PubMedBERT to optimize memory utilization while maintaining computational efficiency. Third, 768-dimensional PubMedBERT embeddings were extracted for each document, capturing semantic relationships between clinical concepts while preserving contextual information. The implementation utilized the SentenceTransformer framework, which provides efficient inference utilities for transformer-based models (detailed in the Appendix E). Given the unsupervised nature of clustering, stability-based validation was employed through random subsampling, feature perturbation, and parameter sensitivity analysis. The complete hardware specifications, details of the software environment and computational optimizations are documented in Appendix D, supporting transparency and reproducibility of the experimental setup.

6.7 Results

The performance of the latent semantic projection framework was evaluated through concordance analysis and label density analysis. These evaluations provide evidence of the model's ability to move beyond simple keyword matching to capture clinically meaningful semantic patterns. To ensure the robustness and generalizability of our results, all supervised classification evaluations were conducted using stratified 10-fold cross-validation. In this approach, the dataset was randomly partitioned into ten approximately equal-sized subsets, with each fold serving as the test set exactly once while the remaining nine were used for training. Performance metrics, including mean accuracy and standard deviation, were calculated across all ten iterations. All reported metrics are averaged across all 10 folds unless otherwise specified.

6.7.1 Validation Against Weak Labels

To evaluate the internal consistency and reliability of the proposed silver standard labels, we conducted a comparative concordance analysis. This phase transitions from unsupervised discovery to computational validation, testing whether the refined categories possess enough semantic distinctiveness to be recovered by supervised classifiers.

Labeling Evolution and Model Performance Improvements in classification accuracy provide empirical support for the effectiveness of the taxonomy refinement. As shown in Table 6.3, the original dictionary-based labels were highly overlapping and noisy, resulting in lower accuracy. In contrast, the refined silver standard achieved an accuracy of 0.92 using a TF-IDF baseline and 0.89 using the PubMedBERT transformer. The marginal lead of the TF-IDF baseline in single-label accuracy suggests that the silver standard retains strong

alignment with the refined vocabulary representation, while the transformer's performance is consistent with the semantic structure of the labels.

Label Source	Classifier	Mean Accuracy (±SD)	Support
Primary Dictionary	TF-IDF + LogReg	0.79 ± 0.03	33,050
	PubMedBERT	0.62 ± 0.04	33,050
Silver Standard	TF-IDF + LogReg	**0.92 ± 0.01**	33,050
	PubMedBERT	**0.89 ± 0.02**	33,050

Table 6.3: Taxonomy Validation: Accuracy (Mean ± SD) Across Labeling Generations

Per-Class Analysis of the Silver Standard Table 6.4 details the performance metrics for the refined super-specialties. The high F1-scores across all domains ranging from 0.81 in Genetics-Endocrine to 0.96 in Ophthalmology demonstrate that the refinement process substantially reduced previous clinical ambiguities.

Domain	Accuracy	Precision	Recall	Specificity	F1-score	Support
Ophthalmology	0.98 ± 0.01	0.95 ± 0.01	0.97 ± 0.01	0.99 ± 0.01	0.96 ± 0.01	1,489
Cardiovascular	0.97 ± 0.01	0.95 ± 0.01	0.94 ± 0.01	0.99 ± 0.01	0.95 ± 0.01	3,992
Oncology-Pathology	0.97 ± 0.01	0.93 ± 0.02	0.94 ± 0.01	0.99 ± 0.01	0.94 ± 0.01	6,114
Musculoskeletal	0.98 ± 0.01	0.93 ± 0.01	0.95 ± 0.01	0.99 ± 0.01	0.94 ± 0.01	4,308
Neurological	0.97 ± 0.01	0.92 ± 0.02	0.91 ± 0.02	0.98 ± 0.01	0.92 ± 0.01	4,352
Gastrointestinal	0.97 ± 0.01	0.92 ± 0.01	0.93 ± 0.02	0.99 ± 0.01	0.92 ± 0.01	3,934
Respiratory	0.97 ± 0.01	0.91 ± 0.02	0.92 ± 0.02	0.98 ± 0.01	0.91 ± 0.02	2,018
Dermatology-Immune	0.96 ± 0.01	0.91 ± 0.02	0.89 ± 0.03	0.98 ± 0.01	0.90 ± 0.02	1,745
Renal-Metabolic	0.96 ± 0.01	0.89 ± 0.02	0.90 ± 0.02	0.98 ± 0.01	0.89 ± 0.02	2,022
Genetics-Endocrine	0.95 ± 0.02	0.92 ± 0.03	0.85 ± 0.04	0.98 ± 0.01	0.88 ± 0.03	908
Infectious-Disease	0.96 ± 0.01	0.89 ± 0.02	0.86 ± 0.03	0.98 ± 0.01	0.87 ± 0.02	2,168
Weighted Average	**0.97 ± 0.01**	**0.92 ± 0.01**	**0.92 ± 0.01**	**0.99 ± 0.01**	**0.92 ± 0.01**	**33,050**

Table 6.4: Classification Report: Silver Standard Validation (Mean±SD, 10-fold CV)

Table 6.5 presents a direct comparison of per-class F1-scores before and after taxonomy refinement. Substantial improvements are observed across all domains, particularly for previously fragmented or low-performing categories.

Domains such as *Renal-Metabolic* and *Infectious-Disease* exhibit the highest gains (+0.42 and +0.55, respectively), indicating that the refined taxonomy effectively consolidates se-

mantically related subdomains into more coherent classes. Although already strong domains such as *Ophthalmology* show modest gains (+0.05), the general trend demonstrates that the proposed taxonomy improves class separability and label consistency across the evaluated medical domains.

Super-specialty Domain	**F1-score (before)**	**F1-score (after)**	**Δ Improvement**
Ophthalmology	0.91	0.96	+0.05
Cardiovascular	0.76	0.95	+0.19
Oncology-Pathology	0.77	0.94	+0.17
Musculoskeletal	0.75	0.94	+0.19
Neurological	0.75	0.92	+0.17
Gastrointestinal	0.76	0.92	+0.16
Respiratory	0.82	0.91	+0.09
Dermatology-Immune	0.57	0.90	+0.33
Renal-Metabolic	0.47	0.89	+0.42
Genetics-Endocrine	0.74	0.88	+0.14
Infectious-Disease	0.32	0.87	+0.55

Table 6.5: Per-Class Performance Improvement After Taxonomy Refinement (F1-score)

Semantic Migration Visualization Figure 6.4 illustrates semantic migration between categories. The heatmap represents the aggregated results across all 10 folds, and reveals that while specialties such as *Ophthalmology* remained pure, there was substantial migration from general labels such as *Surgery* toward more specific pathology labels such as *Oncology*. This suggests that the entity-based model captures underlying clinical context (e.g., *Pathology*) rather than procedural context.

Multi-label Validation: Capturing Clinical Comorbidities Medical records are inherently multi-systemic. To evaluate the proposed taxonomy's ability to handle overlapping medical domains, we conducted a multi-label classification study on the silver standard super-specialties.

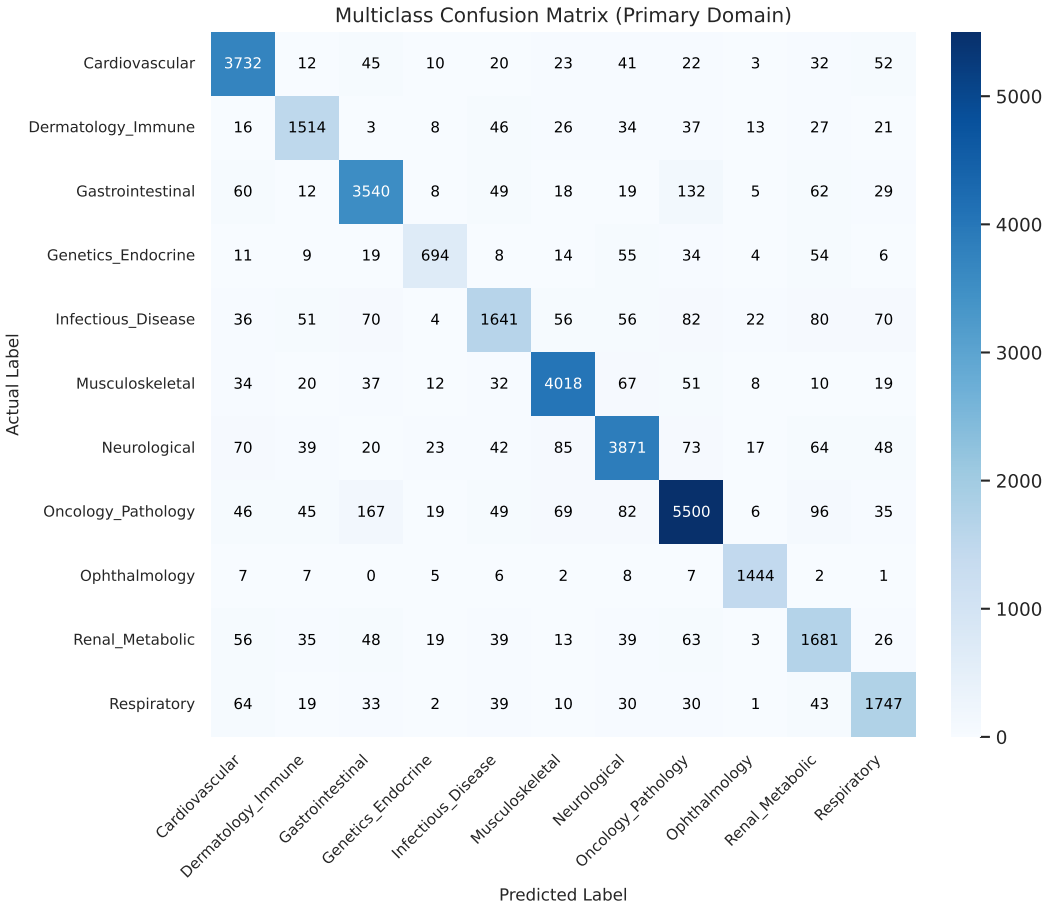

Figure 6.4: Multiclass LSP Silver Standard Confusion Matrix Heatmap.

Architectural Comparison and Model Recall As detailed in Table 6.6, both the TF-IDF baseline and the PubMedBERT transformer achieved a high F1-score of 0.91. However, an analysis of performance metrics reveals a critical distinction: PubMedBERT achieved a substantially higher recall (0.91 vs. 0.87 for the baseline). This suggests that transformer-based models may better capture secondary clinical domains within patient summaries compared to bag-of-words approaches. This gain comes at the cost of reduced precision (0.91 vs. 0.95) and slightly reduced specificity (0.98 vs. 0.99), suggesting that the transformer introduces more false positive label assignments relative to the weak (silver standard) labeling scheme. In the multi-label setting, this behavior is consistent with a model that captures secondary or weaker clinical signals but may over-assign labels relative to the more conservative TF-IDF baseline. Overall, transformer-based models better recover potentially relevant co-occurring specialties, but at the cost of reduced precision and specificity.

The increase in support to $54,567$ in Table 6.6 reflects the transition to a multi-label framework, where the model identifies an average of 1.6 specialties per patient record.

Model	Accuracy	Precision	Recall	Specificity	F1-score	Support
TF-IDF + LogReg	**0.97± 0.01**	**0.95 ± 0.01**	0.87 ± 0.0	**0.99 ± 0.01**	**0.91 ± 0.01**	54,567
PubMedBERT	**0.97± 0.01**	0.91 ± 0.02	**0.91 ± 0.01**	0.98 ± 0.01	**0.91 ± 0.01**	54,567

Table 6.6: Multi-label Validation: PubMedBERT vs. TF-IDF (Mean ± SD).

Semantic Overlap and Co-occurrence Analysis Figure 6.5 illustrates the aggregated label co-occurrence matrix across all 10 cross-validation folds for the 11 super-specialties. The heatmap validates the clinical logic of the silver standard labeling process. Significant overlap is observed between *Oncology_Pathology* and *Gastrointestinal* domains, reflecting the high frequency of oncological presentations within surgical gastrointestinal notes. The ability of the models to maintain a mean F1-score of 0.91 (evaluated via 10-fold cross-validation) despite these semantic overlaps provides evidence that the taxonomy is computationally

effective and clinically representative.

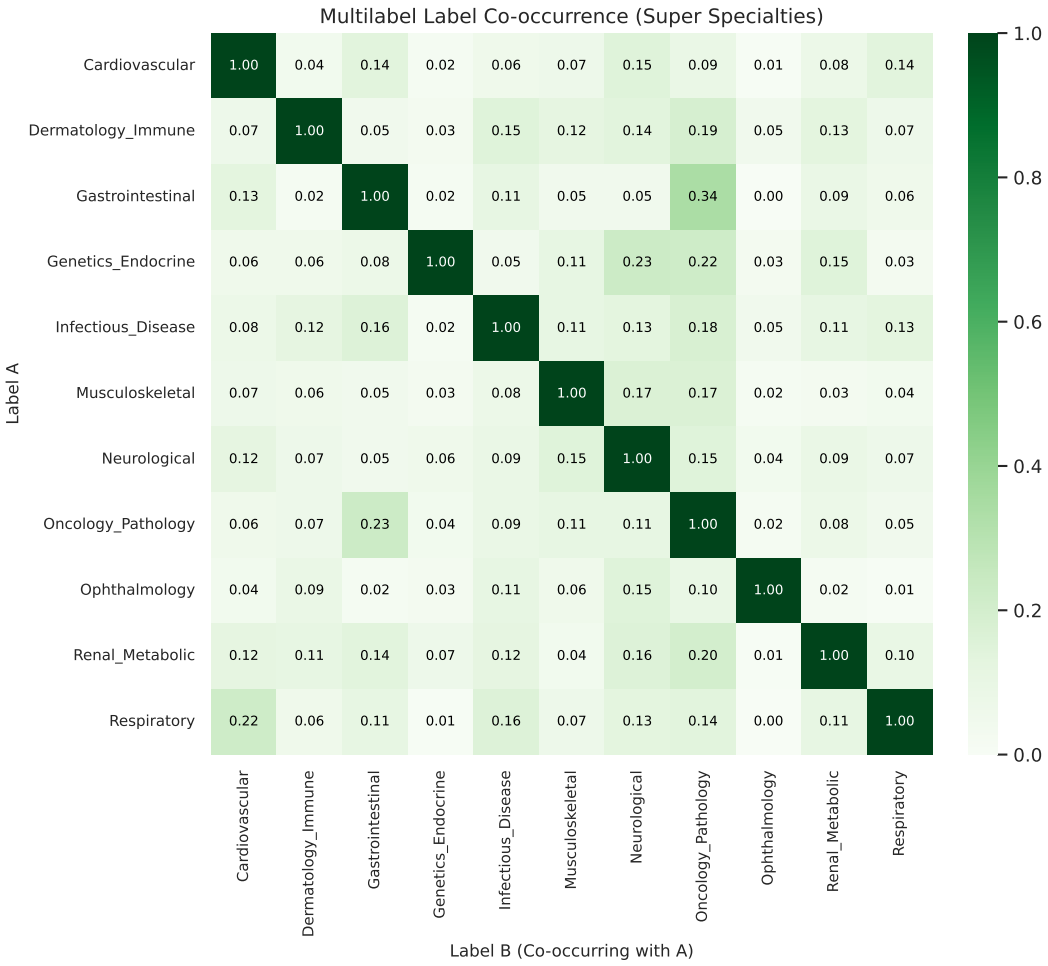

Figure 6.5: Label Co-occurrence Heatmap: Visualizing the Semantic Relationships and Co-morbidity Overlaps Within the Super-specialty Taxonomy

6.7.2 Large Language Model as a Judge

To provide an external qualitative assessment of label coherence, we used a *LLM-as-a-Judge* framework (Li et al., 2024) which uses large language models to categorize patient summaries into specific medical domains. The evaluation was conducted using a lightweight GPT-5-class model **GPT-5 mini** and **GPT-4o**. Each model was provided with the following

system prompt to ensure consistent classification. This evaluation should be interpreted as exploratory qualitative validation rather than a statistically rigorous benchmark.

> *"Role: You are a medical expert. Classify the following patient summaries into one of the following labels: Oncology-Pathology, Neurological, Musculoskeletal, Cardiovascular, Gastrointestinal, Infectious_Disease, Renal_Metabolic, Respiratory, Dermatology_Immune, Ophthalmology, Genetics_Endocrine. Rules: - Only return patient ID and the assigned label. - Do not create new labels; only pick from the provided list."*

Table 6.7 summarizes the performance across 10 samples per super-specialty domain.

Super-specialty	**Samples**	**GPT-4o**	**GPT-5 mini**	**Avg. %**
Ophthalmology	10	10	10	100%
Oncology-Pathology	10	10	8	90%
Genetics-Endocrine	10	8	9	85%
Cardiovascular	10	8	7	75%
Renal-Metabolic	10	8	7	75%
Dermatology-Immune	10	6	8	70%
Infectious-Disease	10	7	7	70%
Respiratory	10	7	6	65%
Neurological	10	6	6	60%
Musculoskeletal	10	6	5	55%
Gastrointestinal	10	4	4	40%

Table 6.7: LLM-as-a-Judge Classification Accuracy

The relatively small sample size (10 instances per super-specialty) was intentionally chosen to enable a high-quality, manual inspection, and controlled evaluation of label coherence across diverse domains. Since the LLM-as-a-Judge framework serves as a qualitative validation rather than a statistical benchmark, the emphasis is on interpretability and error analysis rather than large-scale generalization.

To investigate the nature of the classification errors, Appendix C provides a granular view of the patient records that were misclassified based on results from either GPT-4o or GPT-5 mini, and provides the comprehensive classification breakdown for each patient summary in all 11 medical domains (super-specialties).

The record-level evaluation reveals variability in model performance across medical domains. Although specialties such as *Ophthalmology* achieved 100% agreement within this limited sample, other domains like *Gastrointestinal* (40%) and *Musculoskeletal* ($50 - 60\%$) presented significant challenges for both models.

A critical observation is the high level of inter-model agreement on misclassifications. For instance, in the Gastrointestinal category, both GPT-4o and GPT-5 mini frequently co-identified samples as *Genetics-Endocrine* or *Oncology-Pathology*. Similarly, in the *Respiratory* domain, both models misclassified Patient ID 54648 as *Cardiovascular*. This alignment suggests that the errors may reflect systematic semantic overlap rather than purely random misclassification.

- **Semantic Overlap:** Many patient cases exhibit multi-system involvement, where symptoms of one specialty (e.g., metabolic issues) are clinically intertwined with another (e.g., renal or endocrine systems).

- **Additional hierarchical refinement in the current super-specialty framework:** The current "flat" label list may be insufficient. A hierarchical approach grouping related specialties under broader super-specialty categories before fine-grained labeling could reduce confusion between closely related fields like *Renal-Metabolic* and *Genetics-Endocrine*.

- **Domain Consolidation:** The frequent misclassification of *Gastrointestinal* cases as *Oncology* suggests that for certain datasets, combining these labels or implementing a

multi-label classification strategy may be more representative of clinical reality.

These findings indicate that while LLM-as-a-Judge appears to be a viable evaluation framework, the taxonomy of medical specialties might benefit from further refinement, potentially through the integration of sub-categories to capture the nuance of complex patient summaries.

6.8 Discussion and Chapter Summary

The transition from unsupervised discovery to a supervised validation framework represents a key methodological transition in the methodology of this research. Initial attempts at clinical domain separation using traditional unsupervised representation and clustering approaches resulted in significant semantic overlap. These results provided important diagnostic information, demonstrating that the latent structure of clinical documentation is not fully captured by surface-level lexical co-occurrence but reflects complex clinical intent. By leveraging these insights to engineer a consolidated super-specialty taxonomy, we transformed high-variance, noisy medical tags into a computationally cohesive silver standard.

The subsequent validation of this refined taxonomy produced strong performance. The fact that a traditional TF-IDF baseline achieved a 92% accuracy suggests that the labeling process resulted in distinct lexical patterns associated with each medical domain. Furthermore, the PubMedBERT transformer's performance in the multi-label task achieving a F1-score of 0.91 and a superior recall of 91% provides evidence that the taxonomy is effective in handling the multi-systemic nature of real-world patient records. The alignment of the model's confusion with logical clinical overlaps (e.g., *Oncology* and *Gastrointestinal*) provides qualitative support for the claim that the model captures medical context rather than capturing clinically meaningful semantic relationships rather than relying solely on lexical

memorization.

In conclusion, this chapter demonstrates that, while unsupervised methods alone may be insufficient for clinical classification in isolation, they are an essential precursor to high-quality data engineering. The resulting silver standard dataset and the trained PubMedBERT classifier now provide a foundation for computational support for clinical document organization and triage workflows. By resolving the semantic ambiguities inherent in original hospital dictionary labels, this work suggests a scalable pathway for organizing clinical data into structured, searchable, and clinically meaningful domains.

6.9 Contributions of the Chapter

- **Created interpretable clinical labels**: Silver standard taxonomy reduces semantic ambiguity in clinical narratives.
- **Enable actionable multi-label predictions**: PubMedBERT achieved F1-score of 0.91, high recall, and confusion patterns that align with clinical relationships.
- **Capture contextual relationships within clinical narratives**: Multi-label classification reflects real multimorbidity and clinical intent.
- **Expanded clinically usable structured information**: Multi-label enrichment substantially increases usable information across 165,000 records.
- **Supports efficient and scalable clinical triage**: Transformer-based models demonstrated scalability potential for large clinical corpora for hospital systems.

Chapter 7

Conclusions and Future Work

This chapter summarizes the major outcomes and contributions of the dissertation. It begins with an overview of the primary research contributions and objectives of the work. The chapter then revisits the research questions introduced in Chapter 3 and synthesizes the findings presented throughout the dissertation to demonstrate how each question has been addressed. Next, a detailed discussion of the research contributions is organized according to the four research challenges, highlighting the methodological advances and empirical findings developed in Chapters 4, 5, and 6. The chapter subsequently discusses the broader implications of these findings for medical machine learning, clinical decision-support systems, and healthcare informatics. Finally, the limitations of the proposed approaches are examined, directions for future research are outlined, and concluding remarks are provided regarding the overall significance and potential impact of this work.

7.1 Summary of Research Contributions

This dissertation investigates the evolution of medical data classification, moving from structured tabular analysis to advanced language-based representations. By addressing the pervasive challenges of data scarcity, class imbalance, and categorical ambiguity, this work contributes computational methods that enhance the reliability and scalability of clinical decision-support systems.

We first addressed the limitations of small, unbalanced medical datasets. By evaluating various generative strategies, including conditional tabular generative adversarial networks (CTGANs), forest diffusion, and large language models (LLMs) we showed that synthetic data augmentation can preserve key statistical properties of clinical data while improving model performance. The results suggest that tabular generative approaches can be a viable solution for privacy-constrained medical environments.

Secondly, we introduced a novel tabular-to-text transformation, which pivots from traditional machine learning to natural language processing (NLP). Our comparative analysis showed that while general-purpose models (GPT-4, Llama, Mistral) struggle with specific clinical nuances, domain-specific small language models (SLMs) such as PubMedBERT and BioBERT provide consistently higher accuracy and recall. This finding highlights the critical importance of domain-specific pretraining in medical AI applications.

We finally scaled these insights into an extensive case study on global medical taxonomies. By diagnosing the limitations of unsupervised clustering and designing a silver standard super-specialty taxonomy, we achieved a validation accuracy of 0.92. This provides evidence that the integration of expert-guided taxonomy engineering and transformer-based classification can resolve the semantic ambiguities inherent in hospital records.

7.2 Research Questions and Findings

This section revisits the research questions introduced in Chapter 3 and summarizes how the findings presented throughout this dissertation address each question. The discussion is organized according to the four research challenges identified in Chapter 3.

7.2.1 Challenge 1: The Data Scarcity-Imbalance Paradox

Research Question 1

How do advanced generative models (e.g., GANs, forest diffusion, and LLM-based synthesis) improve classifier performance on the imbalanced UCI cervical cancer dataset?

The experiments presented in Chapter 4 investigated approaches to address the severe class imbalance present in the cervical cancer dataset. Advanced generative methods such as GAN-based approaches, forest diffusion, and LLM-based synthesis are designed to capture more complex feature interactions and produce synthetic records that reflect the underlying data distribution.

The results demonstrate that classifiers trained on augmented dataset using advanced synthetic generation techniques achieved substantial improvements in minority-class detection compared to models trained on the original dataset alone. For example, the random forest classifier increased the recall from 0.16 to 0.46 and ROC-AUC from 0.64 to 0.80. These improvements suggest that advanced generative approaches can provide a mechanism for addressing severe imbalance while preserving clinically meaningful patterns in the data. In general, the findings support the use of modern generative augmentation techniques as an effective strategy to improve classification performance in small and highly imbalanced medical datasets.

Research Question 2

Can generative models produce synthetic tabular patient data that preserves clinically meaningful relationships necessary for robust model training, as evidenced by data distribution analysis and downstream classification performance?

The results of Chapter 4 indicate that synthetic generation methods appear to preserve clinically relevant relationships among patient attributes. Heatmap comparisons between original and generated datasets, together with improvements in downstream classification performance, suggest that synthetic records maintain important statistical patterns present in the original data. The improved diagnostic performance obtained after augmentation further supports the quality and utility of the generated samples for training predictive models.

Research Question 3

What is the optimal integration strategy for combining original and synthetically augmented data to maximize model generalization and minimize overfitting in a low-data regime?

The experiments demonstrate that an effective strategy might be to augment only the training folds with synthetic samples while preserving entirely original data for evaluation. This approach prevents data leakage and ensures that model performance is measured on authentic clinical observations. Across all evaluated classifiers, training with a combination of original and synthetic data consistently outperformed training on the original dataset alone, suggesting that this integration strategy relatively improves generalization in low-data settings.

7.2.2 Challenge 2: The Modality Gap

Research Question 4

Can reformulating structured tabular patient records into natural language descriptions enable language models to capture richer contextual relationships for improved classification?

The results presented in Chapter 5 suggest that transforming tabular records into structured natural language may improve the ability of language models to learn clinically meaningful relationships among risk factors. The proposed tabular-to-text framework allowed language models to process patient information as a coherent clinical narrative rather than as isolated variables. The importance of this representation was further confirmed through ablation studies, where removing the structured transformation resulted in noticeable reductions in classification performance.

Research Question 5

Do domain-specific models (PubMedBERT, BioBERT), pretrained on biomedical corpora, outperform general-purpose models when applied to textual patient representations?

The experimental results suggest the value of domain-specific biomedical pretraining. Among the fine-tuned language models, BioBERT achieved the highest overall performance with an F1-score of 63.73%, outperforming all evaluated prompt-based general-purpose language models. Although few-shot and chain-of-thought prompting improved the performance of general-purpose models, domain-specific models consistently produced superior precision, recall, and F1-scores. These findings highlight the importance of biomedical knowledge embedded during pretraining.

Research Question 6

Does a text-based representation of patient data provide a robust framework for modeling contextual relationships among clinical risk factors?

The findings suggest that text-based representations provide a flexible mechanism for handling heterogeneous clinical information. Feature grouping, normalization, and structured patient descriptions helped language models to capture contextual relationships between risk factors while reducing sensitivity to variations in individual variables. The positive results observed across both cervical cancer classification and large-scale clinical narrative analysis show that language-based representations may support robust clinical modeling across diverse data sources.

7.2.3 Challenge 3: The Deployment Dilemma

Research Question 7

Can data-augmented machine learning models and domain-specific language models provide an effective balance between predictive performance and computational efficiency?

The results of Chapters 4 and 5 demonstrate that predictive performance may be achieved without relying only on extremely large computational models. Data-augmented traditional machine learning models achieved substantial performance improvements, while compact domain-specific language models such as BioBERT showed improvement in classification results using comparatively modest computational resources. Together, these findings support the practicality of combining efficient machine learning methods with targeted NLP approaches for deployment-oriented clinical decision support.

Research Question 8

To what extent can zero-shot and few-shot learning with pretrained language models provide a viable low-data alternative to fully supervised training?

The experiments show that zero-shot, few-shot, and chain-of-thought prompting allow meaningful classification performance even without task-specific fine-tuning. Performance improved as additional prompting structure was introduced, and chain-of-thought prompting showed the strongest results among the evaluated prompting strategies. Although supervised domain-specific models achieved higher overall scores, these findings suggest that prompt-based methods might provide a useful alternative when labeled data are limited.

Research Question 9

How effectively can the proposed framework operate in a low-data medical environment characterized by limited labeled samples and class imbalance?

Across the experiments conducted in this dissertation, the proposed method demonstrated relatively stable performance despite the small size and class imbalance of the cervical cancer dataset. The combination of synthetic data augmentation, domain-specific pretraining, and cross-validation helped the models maintain predictive capability in a low-data environment. These results suggest that our approach appears to be helpful in settings where large annotated medical datasets are not readily available.

7.2.4 Challenge 4: The Interpretability-Actionability Disconnect

Research Question 10

Does the proposed text-based modeling framework offer a more natural pathway toward generating interpretable explanations compared to feature-attribution methods applied to tabular

models?

The results obtained throughout this dissertation suggest that language-based representations can provide a natural foundation for clinically meaningful explanations. Chain-of-thought prompting produced human-readable reasoning patterns that explicitly referenced patient risk factors, while the silver standard taxonomy developed in Chapter 6 organized clinical narratives into interpretable medical domains. Together, these findings demonstrate that text-based approaches naturally appear to align predictive outputs with the language used in clinical reasoning and communication.

This section outlines our primary research contributions, which are detailed further in the following subsections.

7.3 Research Contributions

- **Challenge 1: Data Scarcity–Imbalance Paradox**
  - **Improve classifier performance under severe imbalance and limited data.**

    Method: Synthetic data generation (GANs, forest diffusion, LLM-based synthesis).

    Contribution: The use of synthetic data allows machine learning methods to match or exceed their baseline performance.
  - **Preserve multivariate clinical relationships.**

    Method: Evaluation of statistical similarity and downstream fidelity.

    Contribution: This approach enables synthetic datasets to retain important clinical relationships while improving predictive performance.
  - **Identify optimal integration of synthetic and original data.**

Method: Comparative analysis between synthetic and original datasets.

Contribution: Combining synthetic samples with original training data while preserving authentic evaluation data improves model generalization and provides a practical augmentation strategy for low-data medical settings.

- **Challenge 2: Modality Gap (Structured to Contextual)**
  - **Determine tabular-to-text contextual modeling improvement.**

    Method: Tabular-to-text transformation.

    Contribution: Transforming structured data into textual representations enables language models to capture richer contextual and semantic relationships.
  - **Compare domain-specific vs. general large language models.**

    Method: BioBERT , PubMedBERT, and DistilBERT.

    Contribution: The use of domain-specific models enables superior performance compared to general-purpose LLMs in multiple evaluation metrics.
  - **Model contextual relationships among clinical risk factors.**

    Method: SLM fine-tuning on structured textual patient representations.

    Contribution: Structured textual representations enable language models to capture relationships among clinical risk factors beyond isolated tabular features.
- **Challenge 3: Deployment Dilemma**
  - **Achieve accuracy–efficiency trade-off.**

    Method: Comparative evaluation of SLMs and LLMs.

    Contribution: The use of smaller language models enables competitive recall while significantly reducing computational cost, making them suitable for resource-constrained deployment.

- **Assess viability of zero-shot and few-shot learning.**

  Method: Zero-shot, few-shot, and chain-of-thought prompting.

  Contribution: Prompt-based learning approaches reveal that few-shot and chain-of-thought strategies improve recall, although performance remains variable.

- **Evaluate performance under low-data conditions.**

  Method: Synthetic data augmentation and SLM-based modeling.

  Contribution: The combination of synthetic data and SLMs enables stable and competitive performance even with limited real-world data.

- **Challenge 4: Interpretability–Actionability**

  - **Enable interpretable, clinically meaningful outputs.**

    Method: Chain-of-thought prompting.

    Contribution: The use of chain-of-thought reasoning enables transparent, step-by-step clinical explanations.

  - **Develop a clinical taxonomy for narrative data.**

    Method: Unsupervised clustering to construct a silver standard taxonomy.

    Contribution: This approach enables the creation of a clinically meaningful labeling system that reduces semantic ambiguity.

  - **Validate interpretability through multi-label classification.**

    Method: PubMedBERT multi-label classifier.

    Contribution: The resulting model achieves an F1-score of 0.91, with high recall and confusion patterns that align with real clinical overlaps.

7.3.1 Challenge 1: Contributions to Addressing Data Scarcity and Class Imbalance

Improving Classification Performance in Low-Resource Medical Datasets

A major contribution of this dissertation is the demonstration that synthetic data augmentation can substantially improve classification performance in a small and highly imbalanced medical dataset. As presented in Chapter 4, multiple synthetic data generation techniques were evaluated and integrated with traditional machine learning classifiers. The results showed consistent improvements in recall, F1-score, and ROC-AUC, particularly for the Random Forest classifier. These findings provide evidence that synthetic data augmentation can enhance minority-class detection while preserving strong overall predictive performance.

Preserving Clinically Meaningful Relationships in Synthetic Data

This dissertation also contributes an evaluation framework to assess whether the generated medical data maintain a clinically useful structure. Through statistical comparisons and downstream classification experiments, the generated records demonstrated sufficient fidelity to support reliable model training. The results suggest that synthetic datasets can serve as valuable complements to limited clinical datasets while preserving important relationships among risk factors.

Establishing an Effective Augmentation Strategy

A further contribution is the identification of an effective strategy for integrating synthetic and original data. By generating synthetic samples only within the training folds and maintaining fully authentic data for evaluation, the proposed methodology avoids data leakage while improving model generalization. This experimental design provides a reproducible

framework for future studies involving medical data augmentation.

7.3.2 Challenge 2: Contributions to Bridging the Modality Gap

Development of a Tabular-to-Text Transformation Framework

One of the central methodological contributions of this dissertation is the development of a structured tabular-to-text transformation described in Chapter 5. This approach converts structured patient records into coherent natural-language descriptions that can be processed by modern language models. This enables clinical risk factors to be represented as contextual narratives rather than isolated variables.

Demonstrating the Value of Domain-Specific Language Models

This research provides a systematic comparison between general-purpose language models and domain-specific biomedical language models. The results show that models such as BioBERT and PubMedBERT consistently outperform general-purpose systems across multiple evaluation metrics. These findings reinforce the importance of biomedical pretraining and contribute empirical evidence supporting domain-adapted language models for healthcare applications.

Improving Robustness Through Language-Based Representations

The dissertation further demonstrates that structured textual representations support robust clinical classification. Feature grouping, normalization, and contextual reformulation enable models to capture relationships among risk factors more effectively than traditional feature-based approaches. The supporting ablation studies presented in Chapter 5 confirm the importance of these design decisions.

7.3.3 Challenge 3: Contributions to Practical Deployment

Balancing Predictive Performance and Computational Efficiency

This dissertation contributes a deployment-oriented perspective on medical AI by emphasizing efficient and practical solutions rather than relying exclusively on increasingly large models. The results demonstrate that data-augmented machine learning methods and compact domain-specific language models can achieve strong predictive performance while maintaining a computational footprint suitable for real-world healthcare environments.

Evaluating Prompt-Based Learning Under Limited Data Conditions

Another contribution is the evaluation of zero-shot, few-shot, and chain-of-thought prompting strategies for medical classification. The experiments showed that prompt engineering can meaningfully improve the performance of general-purpose models and provide useful alternatives when access to labeled data is restricted.

Supporting Clinical Modeling in Data-Constrained Environments

The combined use of synthetic data generation, transfer learning, and domain-specific language models provides a framework specifically designed for low-data clinical environments. The results demonstrate that competitive performance can be achieved without requiring large proprietary datasets, increasing the potential accessibility of AI-based decision-support systems.

7.3.4 Challenge 4: Contributions to Interpretability and Clinical Actionability

Supporting Explainable Clinical Reasoning

This dissertation contributes evidence that language-based modeling approaches naturally support interpretable outputs. Unlike traditional tabular predictions that often require post-hoc explanation methods, text-based approaches can represent clinical reasoning directly through natural language descriptions and structured reasoning processes.

Development of the Silver Standard Clinical Taxonomy

A major contribution of Chapter 6 is the development of a silver standard taxonomy for organizing large collections of clinical narratives. The methodology transformed noisy and overlapping medical labels into a clinically meaningful set of super-specialties, producing a structured representation suitable for large-scale analysis and classification.

Validation of Clinically Meaningful Multi-Label Classification

The final contribution is the demonstration that transformer-based models can successfully classify clinical records using this refined taxonomy. The resulting multi-label framework achieved strong performance while preserving clinically meaningful specialty relationships and comorbidities. These findings support the broader goal of developing decision-support systems that align more closely with real clinical workflows.

7.4 Discussion of Findings

Several key themes emerged across the different stages of this research. First, the transition from tabular to text-based modeling represented more than a simple change in format; it allowed for a richer, and more contextual representation of patient records. This was particularly evident in the final case study, where the model's ability to handle multi-label comorbidities (F1-score: 0.91).

Second, the limitations of unsupervised methods (LDA-PCA) in the early stages of the case study provided a crucial meta-finding: clinical data pose challenges for simple word-frequency-based methods. The success of the subsequent supervised PubMedBERT model suggests that deep learning architectures are not limited to keyword matching, but are capable of learning clinical intent, provided they are supported by a logically structured taxonomy.

Finally, this work underscores the value of the silver standard approach. By refining messy, high-variance primary specialties into cohesive super-specialties, we demonstrated that data engineering is as vital as model architecture. Substantial improvement in performance serves as strong empirical evidence for the use of transformers in large-scale clinical evaluation.

7.5 Limitations

Despite promising results, this research is not without limitations. The generative models used in this study, while effective, require careful post-processing to ensure that synthetic samples do not introduce clinical hallucinations or impossible feature combinations.

Furthermore, the silver standard taxonomy, while computationally robust, still relies on the quality of the initial clinical summaries. In cases where documentation is extremely brief

or lacks specific anatomical terminology, even domain-specific models like PubMedBERT reach a performance ceiling. Lastly, the multi-label overlaps observed in the final chapter while clinically logical suggest that a small percentage of medical records will always remain ambiguous due to the inherent complexity of multi-systemic diseases.

7.6 Future Work

Building upon the foundations laid in this dissertation, several avenues for future research remain open:

1. **Dynamic Taxonomy Expansion:** Future studies could explore the use of hierarchical classification to expand the 11 super-specialties back into more granular sub-specialties. This would allow for a triage-to-treatment pipeline where a record is first categorized into a major domain and then sub-classified for specific clinical pathways.

2. **Clinical Decision Validation:** A critical next step is the prospective validation of these models in a real-world clinical setting. Collaborating with healthcare providers to compare model-suggested categories against real-time physician decisions would provide a critical evaluation of the system's clinical applicability.

3. **Multimodal Integration:** Integrating structured tabular data (vitals, lab results) directly with text-based summaries (PubMedBERT embeddings) could create a multimodal representation of the patient. This hybrid approach may further enhance the accuracy of cervical cancer risk assessments and general domain classification.

4. **Dialogue-Based Decision Support:** Finally, the success of text-based modeling suggests the potential for a dialog-based interface. This system could communicate

directly with patients or clinicians in natural language, guiding them to the necessary follow-up tests based on the classification of their clinical history.

7.7 Concluding Remarks

This dissertation suggests that the future of medical AI may increasingly depend on the synergy between structured data modeling and language-based context. By transforming noisy, sparse and ambiguous medical records into a refined silver standard and applying domain-specific deep learning, we can bridge the gap between dark clinical data and actionable medical insights. The methodologies developed in this dissertation outline a pathway to an extensible framework for clinical decision-support systems, capable of evolving with new data sources, domain knowledge, and modeling paradigms.

Bibliography

Sherif F. Abdoh, Mohamed A. Rizka, and Fahima A. Maghraby. Cervical cancer diagnosis using random forest classifier with SMOTE and feature reduction techniques. *IEEE Access*, 6:59475–59485, 2018.

Naji R. Aljohani, Dhabiya Al-Thani, Maryam Al-Mansoori, Hessa AlMarri, Fatima AlMarzooqi, Maryam Al-Kuwari, and Khulood Al-Thani. Optimizing patient stratification using electronic health records and machine learning: A systematic review. *Journal of Healthcare Informatics Research*, 8(1):1–27, 2024.

Turki Aljrees. Improving prediction of cervical cancer using KNN imputer and multi-model ensemble learning. *PLOS ONE*, 19(1):e0295632, 2024.

Farah Alshanik, Amy Apon, Alexander Herzog, Ilya Safro, and Justin Sybrandt. Accelerating text mining using domain-specific stop word lists. In *IEEE International Conference on Big Data*, pages 2639–2648, 2020.

Naomi S. Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992. doi: 10.1080/00031305.1992.10475879.

American Board of Medical Specialties. American Board of Medical Specialties (ABMS), 2026. URL `https://www.abms.org/`. Accessed: 2025-04-29.

Marc Arbyn, Elisabete Weiderpass, Laia Bruni, Silvia de Sanjose, Mona Saraiya, Jacques Ferlay, and Freddie Bray. Estimates of incidence and mortality of cervical cancer in 2018: A worldwide analysis. *The Lancet Global Health*, 8(2):e191–e203, 2020.

Robert Arnold, Amish Shah, Ramesh Karkal, Vikas Kumar, Rameez Kazi, Manjunath Madhira, Sridhar Saravanan, Kannan Ramamurthy, and Pavan Kumar. Topic modeling for qualitative analysis of scientific literature. *Journal of Biomedical Informatics*, 43(5):747–757, 2010.

Michael Bamberg. Language, concepts and emotions: The role of language in the construction of emotions. *Language Sciences*, 19(4):309–340, 1997.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

Verónica Bolón-Canedo and Amparo Alonso-Betanzos. *Feature Selection for High-Dimensional Data*, volume 16 of *Studies in Big Data*. Springer, Cham, 2015. doi: 10.1007/978-3-319-21858-8.

Vadim Borisov, Karl Sebler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. Language models are realistic tabular data generators. In *International Conference on Learning Representations (ICLR)*, 2023. URL `https://openreview.net/forum?id=cEygmQNOeI`.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Bagaria Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Max Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

Fan Cao, Yi-Zi Li, De-Yu Zhang, Xiao-Ying Wang, Wen-Xiao Chen, Fang-Hua Liu, Yi-Xuan Men, Song Gao, Chun-Qing Lin, and Hua-Chun Zou. Human papillomavirus infection and the risk of cancer at specific sites. *eBioMedicine*, 104, 2024.

Davide Chicco, Andrea Campagner, Alice Spagnolo, Davide Ciucci, and Giuseppe Jurman. The Silhouette coefficient and the Davies-Bouldin index are more informative than Dunn index, Calinski-Harabasz index, Shannon entropy, and Gap statistic for unsupervised clustering internal evaluation of two convex clusters. *PeerJ Computer Science*, 11:e3309, 2025.

Avishek Choudhury, Y. M. Wesabi, and Daehan Won. Classification of cervical cancer dataset. *arXiv preprint arXiv:1812.10383*, 2018.

Dearbhaile C. Collins, Raghav Sundar, Joline S. J. Lim, and Timothy A. Yap. Towards precision medicine in the clinic: From biomarker discovery to novel therapeutics. *Trends in Pharmacological Sciences*, 38(1):25–40, 2017.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3): 273–297, 1995. doi: 10.1007/BF00994018.

Lynette Denny, Silvia de Sanjose, Miriam Mutebi, Benjamin O. Anderson, Jane Kim, Jose Jeronimo, Rolando Herrero, Karen Yeates, Ophira Ginsburg, and Rengaswamy Sankaranarayanan. Interventions to close the divide for women with breast and cervical cancer between low-income and middle-income countries and high-income countries. *The Lancet*, 389(10071):861–870, 2017.

Juan G Diaz Ochoa, Natalie Layer, Jonas Mahr, Faizan E Mustafa, Christian U Menzel, Martina Müller, Tobias Schilling, Gerald Illerhaus, Markus Knott, and Alexander Krohn. Optimized BERT-based nlp outperforms zero-shot methods for automated symptom detection in clinical practice. *Frontiers in Digital Health*, 7:1623922, 2025.

Rabie Adel El Arab, Mohammad S Abu-Mahfouz, Fuad H Abuadas, Husam Alzghoul, Mohammed Almari, Ahmad Ghannam, and Mohamed Mahmoud Seweid. Bridging the gap: from AI success in clinical trials to real-world healthcare implementation—a narrative review. In *Healthcare*, volume 13, page 701, 2025.

Naeem Fatima, Hafiz Muhammad Afzaal, Muhammad Khalid Mehmood Hussain, and Muhammad Sajid. Language and emotion: A study of emotional expression in multilinguals. *Journal of Applied Linguistics and TESOL*, 7(4):932–946, 2024.

Salma Fayaz, Syed Zubair Ahmad Shah, Nusrat Mohi ud din, Naillah Gul, and Assif Assad. Advancements in data augmentation and transfer learning: A comprehensive survey to address data scarcity challenges. *Recent Advances in Computer Science and Communications*, 17(8):14–35, 2024.

Kelwin Fernandes, Jaime Cardoso, and Jessica Fernandes. Cervical cancer (risk factors), 2017. UCI Machine Learning Repository.

Maayan Frid-Adar, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Synthetic data augmentation using GAN for improved liver lesion classification. *IEEE Transactions on Medical Imaging*, 38(3):834–843, 2018. doi: 10.1109/TMI.2018.2865672.

Carol Friedman, Lyudmila Shagina, Yves Lussier, and George Hripcsak. Automated encoding of clinical documents based on natural language processing. *Journal of the American Medical Informatics Association*, 11(5):392–402, 2004.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001. doi: 10.1214/aos/1013203451.

Sunyang Fu, David Chen, Huan He, Sijia Liu, Sungrim Moon, Kevin J Peterson, Feichen Shen, Liwei Wang, Yanshan Wang, Andrew Wen, Yiqing Zhao, Sunghwan Sohn, and Hongfang Liu. Clinical concept extraction: a methodology review. *Journal of Biomedical Informatics*, 109:103526, 2020.

Laya Elsa George and Lokendra Birla. A study of topic modeling methods. In *IEEE International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 109–113, 2018.

Reza Gheibi and Dean Hougen. Enhancing cervical cancer classification through augmented and synthetic data in machine learning. In *IEEE International Conference on Prognostics and Health Management (ICPHM)*, pages 1–8. IEEE, 2025a.

Reza Gheibi and Dean Hougen. Improved cervical cancer classification using generative ai and language models. In *IEEE International Conference on Prognostics and Health Management (ICPHM)*, pages 1–8. IEEE, 2025b.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2014.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. PubMedBERT: A pretrained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:2007.15748*, 2020a.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. PubMedBERT: Domain-specific language model pretraining for biomedical natural language processing. *arXiv preprint arXiv:2007.15779*, 2020b.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, 2021.

Fatih Gurcan and Ahmet Soylu. Learning from imbalanced data: integration of advanced resampling techniques and machine learning models for enhanced cancer diagnosis and prognosis. *Cancers*, 16(19):3417, 2024.

Harvard Medical School. How artificial intelligence is disrupting medicine and what it means for physicians, 2023. URL `https://postgraduateeducation.hms.harvard.edu`.

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. TabLLM: Few-shot classification of tabular data with large language models. In *International Conference on Artificial Intelligence and Statistics*, pages 5549–5581. PMLR, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

John H. Holmes, James Beinlich, Mary R. Boland, Kathryn H. Bowles, Yong Chen, Tessa S. Cook, George Demiris, Michael Draugelis, Laura Fluharty, Peter E. Gabriel, Robert Grundmeier, C. William Hanson, Daniel S. Herman, Blanca E. Himes, Rebecca A. Hubbard, Charles E. Kahn Jr., Dokyoon Kim, Ross Koppel, Qi Long, Nebojsa Mirkovic, Jeffrey S. Morris, Danielle L. Mowery, Marylyn D. Ritchie, Ryan Urbanowicz, and Jason H. Moore. Why is the electronic health record so challenging for research and clinical care? *Methods of Information in Medicine*, 60:032–048, 2021.

David W. Hosmer, Stanley Lemeshow, and Rodney X. Sturdivant. *Applied Logistic Regression*. John Wiley & Sons, 3rd edition, 2013. doi: 10.1002/9781118548387.

Elias Hossain, Rajib Rana, Niall Higgins, Jeffrey Soar, Prabal Datta Barua, Anthony R. Pisani, and Kathryn Turner. Natural language processing in electronic health records in relation to healthcare decision-making: A systematic review. *Computers in Biology and Medicine*, 155:106649, 2023.

Xin Hou, Guangyang Shen, Liqiang Zhou, Yinuo Li, Tian Wang, and Xiangyi Ma. Artificial intelligence in cervical cancer screening and diagnosis. *Frontiers in Oncology*, 12:851367, 2022.

Liming Hu, David Bell, Sameer Antani, Zhiyun Xue, Kai Yu, Matthew P. Horning, Noni Gachuhi, Benjamin Wilson, Mayoore S. Jaiswal, Brian Befano, L. Rodney Long, Rolando Herrero, Mark H. Einstein, Robert D. Burk, Maria Demarco, Julia C. Gage, Ana Cecilia Rodriguez, Nicolas Wentzensen, and Mark Schiffman. An observational study of deep learning and automated evaluation of cervical images for cancer screening. *JNCI: Journal of the National Cancer Institute*, 111(9):923–932, 2019.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1): 193–218, 1985.

Hugging Face. Trainer API documentation, 2024. URL `https://huggingface.co/docs/transformers/main_classes/trainer`. Accessed: 2025-04-29.

Fei Jiang, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, Yilong Wang, Qiang Dong, Haipeng Shen, and Yongjun Wang. Artificial intelligence in healthcare: Past, present and future. *Stroke and Vascular Neurology*, 2(4), 2017.

Justin M. Johnson and Taghi M. Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54, 2019.

Alexia Jolicoeur-Martineau, Kilian Fatras, and Tal Kachman. Generating and imputing tabular data via diffusion and flow-based gradient-boosted trees. *https://arxiv.org/abs/2309.09968*, 2023. URL `https://arxiv.org/abs/2309.09968`.

Mihir Kale and Abhinav Rastogi. Text-to-text pre-training for data-to-text tasks. *arXiv preprint arXiv:2005.10433*, 2020.

Won Kim, Lana Yeganova, Donald C. Comeau, and W. John Wilbur. Identifying well-formed biomedical phrases in MEDLINE® text. *Journal of Biomedical Informatics*, 45 (6):1035–1041, 2012.

Theresa A Koleck, Caitlin Dreisbach, Philip E Bourne, and Suzanne Bakken. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. *Journal of the American Medical Informatics Association*, 26:364–379, 2019.

David Krech, Richard S. Crutchfield, and Egerton L. Ballachey. *Individual in Society: A Textbook of Social Psychology*. McGraw-Hill, New York, 1962.

Anjali Kuruvilla and B. Jayanthi. Ensemble feature selection and deep learning ensemble classifier for cervical cancer diagnosis. *International Journal of Bioinformatics Research and Applications*, 19(5–6):430–461, 2023.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019a. URL `https://arxiv.org/abs/1909.11942`.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019b.

Thomas K. Landauer, Peter W. Foltz, and Darrell Laham. *Introduction to Latent Semantic Analysis*. HCI International, 1998.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521:436–444, 2015.

Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. LLMs-as-judges: a comprehensive survey on LLM-based evaluation methods. *arXiv preprint arXiv:2412.05579*, 2024.

Xiaolong Li, Yanyan Chen, Peng Zhang, and Xiaofei Zhang. Hybrid approach for sepsis detection using electronic health records and wearable sensor data. *IEEE Journal of Biomedical and Health Informatics*, 23(6):2361–2369, 2019.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sanchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

Christian Lopez, Scott Tucker, Tarik Salameh, and Conrad Tucker. An unsupervised machine learning method for discovering patient clusters based on genetic signatures. *Journal of Biomedical Informatics*, 85:30–39, 2018.

Jiayi Lu, Enmin Song, Ahmed Ghoneim, and Mubarak Alrashoud. Machine learning for assisting cervical cancer diagnosis: An ensemble approach. *Future Generation Computer Systems*, 106:199–205, 2020.

Masoud Makrehchi and Mohamed S. Kamel. Automatic extraction of domain-specific stopwords from labeled documents. In *European Conference on Information Retrieval*, pages 222–233. Springer, 2008.

Ggaliwango Marvin, Hellen Nakayiza, Daudi Jjingo, and Joyce Nakatumba-Nabende. Prompt engineering in large language models. In *International Conference on Data Intelligence and Cognitive Informatics*, pages 387–402. Springer, 2023.

Sai Srinivas Matta and Manish Bolli. Federated learning for privacy-preserving healthcare data sharing: Enabling global AI collaboration. *American Journal of Scholarly Research and Innovation*, 4(01):320–351, 2025.

Muhammad Mehmood, Muhammad Rizwan, Martin Gregus ml, and Sajid Abbas. Machine learning assisted cervical cancer detection. *Frontiers in Public Health*, 9:788376, 2021. doi: 10.3389/fpubh.2021.788376.

Meta LLama Team. LLaMA models on Hugging Face, 2024. URL `https://huggingface.co/meta-llama/Llama-3.2-1B`. Accessed: 2025-05-25.

Stephane M. Meystre, Guergana K. Savova, Karin C. Kipper-Schuler, and John F. Hurdle. Extracting information from textual electronic health records: A review of recent characteristics. *Yearbook of Medical Informatics*, 17(1):128–144, 2008.

Microsoft. Phi-4 technical report, 2024. URL `https://huggingface.co/microsoft/phi-4`. Accessed: 2025-05-25.

Mistral AI. Mistral-7b-v0.1 on Hugging Face, 2023. URL `https://huggingface.co/mistralai/Mistral-7B-v0.1`. Accessed: 2025-05-25.

Muadalina Maria Muraru, Zsuzsa Simo, and Laszlo Barna Iantovics. Cervical cancer prediction based on imbalanced data using machine learning algorithms. *Applied Sciences*, 14 (22):10085, 2024.

Kevin P. Murphy. Naive Bayes classifiers. *University of British Columbia Technical Report*, 18(60):1–8, 2006.

Ritu Nayar and David C. Wilbur. *The Bethesda System for Reporting Cervical Cytology: Definitions, Criteria, and Explanatory Notes*. Springer, 2015.

Emonena Patrick Obrik-Uloho, Oluwaseun Oladeji Olaniyi, Olubukola Omolara Adebiyi, Rukayat Oluwabukola Olasege, and Seun Michael Oyekunle. Dark data in digital health: A predictive framework for identifying and utilizing underreported clinical signals. *Asian Journal of Research in Computer Science*, 18(11):60–74, 2025.

Emmanuel O. Oluwole, A. S. Mohammed, M. R. Akinyinka, and Olufunmilayo Salako. Cervical cancer awareness and screening uptake among rural women in Lagos, Nigeria. *Journal of Community Medicine and Primary Health Care*, 29(1):81–88, 2017.

O. B. Omowhara, S. Ameh, and A. A. Banjo. Cervical cancer screening awareness and uptake among under-screened women in a rural Nigerian community. *Nigerian Health Journal*, 22(4):428–432, 2022.

Nir Oren. Re-examining TF-IDF-based information retrieval with genetic programming. In *Proceedings of SAICSIT*, pages 1–10, 2002.

Rafael M. Pereira, David R. Bertolini, Lucas O. Teixeira, Carlos N. Silla, and Yandre M. Costa. COVID-19 identification in chest X-ray images on flat and hierarchical classification scenarios. *Computer Methods and Programs in Biomedicine*, 194:105532, 2020. doi: 10.1016/j.cmpb.2020.105532.

Nicolas Perotte, Robert J. Tibshirani, and Trevor Hastie. Diagnosis-driven topic modeling for electronic health records. *Journal of the American Medical Informatics Association*, 21(1):119–126, 2014.

Niki Z. Petrakos, Erica E. M. Moodie, and Nicolas Savy. A framework for generating realistic synthetic tabular data in a randomized controlled trial setting. *Statistics in Medicine*, 44 (18–19):e70227, 2025.

Tra My Pham, Nikolaos Pandis, and Ian R. White. Missing data: Issues, concepts, methods. In *Seminars in Orthodontics*, volume 30, pages 37–44. Elsevier, 2024.

W. Nicholson Price and I. Glenn Cohen. Privacy in the age of medical big data. *Nature Medicine*, 25(1):37–43, 2019.

J. Ross Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986. doi: 10.1007/BF00116251.

Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M. Dai, Nissan Hajaj, Michaela Hardt, Peter J. Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, Patrik Sundberg, Hector Yee, Kun Zhang, Yi Zhang, Gerardo Flores, Gavin E. Duggan, Jamie Irvine, Quoc Le, Kurt Litsch, Alexander splinter Mossin, Justin Tansuwan, De Wang, James Wexler, Jimbo Wilson, Dana Ludwig, Samuel L. Volchenboum, Katherine Chou, Michael Pearson, Srinivasan Madabushi, Nigam H. Shah, Atul J. Butte, Michael D. Howell, Claire Cui, Greg S. Corrado, and Jeffrey Dean. Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 1:18, 2018.

Juan Ramos. Using TF-IDF to determine word relevance in document queries. In *Proceedings of the First Instructional Conference on Machine Learning*, pages 29–48, 2003.

Dietrich Rebholz-Schuhmann, Antonio José Jimeno Yepes, Erik M Van Mulligen, Ning Kang, Jan Kors, David Milward, Peter Corbett, Ekaterina Buyko, Elena Beisswanger, and Udo Hahn. CALBC silver standard corpus. *Journal of Bioinformatics and Computational Biology*, 8(01):163–179, 2010.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT networks. *arXiv preprint arXiv:1908.10084*, 2019.

Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986. doi: 10.1038/323533a0.

Ahmed M Salih, Zahra Raisi-Estabragh, Ilaria Boscolo Galazzo, Petia Radeva, Steffen E Petersen, Karim Lekadir, and Gloria Menegaz. A perspective on explainable artificial intelligence methods: SHAP and LIME. *Advanced Intelligent Systems*, 7:2400304, 2025.

Mabrouka Salmi, Dalia Atif, Diego Oliva, Ajith Abraham, and Sebastian Ventura. Handling imbalanced medical datasets: review of a decade of research. *Artificial Intelligence Review*, 57(10):273, 2024.

Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523, 1988.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT: A distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019a.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019b.

Edward Sapir. *Language: An Introduction to the Study of Speech*. Harcourt, Brace & Sons, New York, 1921.

Tom M. Seinen, Jan A. Kors, Erik M. van Mulligen, and Peter R. Rijnbeek. Using structured codes and free-text notes to measure information complementarity in electronic health records: Feasibility and validation study. *Journal of Medical Internet Research*, 27:e66910, 2025.

Seyedmostafa Sheikhalishahi, Riccardo Miotto, Joel T. Dudley, Alberto Lavelli, Fabio Rinaldi, and Azeddine Osmani. Natural language processing of clinical notes on chronic diseases: Systematic review. *JMIR Medical Informatics*, 8(4):e12264, 2020.

Michelle B. Shin, Gui Liu, Nelly Mugo, and Rafael Garcia. A framework for cervical cancer elimination in low- and middle-income countries. *Frontiers in Public Health*, 9:670032, 2021.

Connor Shorten and Taghi M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, 2019. doi: 10.1186/s40537-019-0197-0.

Yi Tay, Wonseok Chung, and Lei Xie. Shorter prompts lead to more efficient language model processing. *arXiv preprint arXiv:2012.15832*, 2020. URL `https://arxiv.org/abs/2012.15832`.

Catriona Tullo and James R. Hurford. Modelling Zipfian distributions in language, 2003. Technical Report, University of Edinburgh.

G Vanitha and M Kasthuri. Enhancing classification of high-dimensional neurocognitive disorder data via neural network-based feature selection with static attention and noise robustness. *Indian Journal of Science and Technology*, 18(36):2964–2974, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Leon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4454–4457. IEEE, 2010.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11:2837–2874, 2010.

Haoran Wang, Qiuye Jin, Shiman Li, Siyu Liu, Manning Wang, and Zhijian Song. A comprehensive survey on deep active learning in medical image analysis. *Medical Image Analysis*, 95:103201, 2024.

Ke Wang, Sungjun Cho, and Ji Wu. Tabular to text: A framework for training language models on structured data. *https://arxiv.org/abs/2210.06280*, 2022. URL `https://arxiv.org/abs/2210.06280`.

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, and Hongfang Liu. Clinical natural language processing in the electronic health records era: A review. *Journal of the American Medical Informatics Association*, 25(6):740–744, 2018a.

Yifan Wang, Wei Zhang, and Jun Zhao. Clinical notes and structured EHR data for mortality prediction: A comparative study. *Journal of the American Medical Informatics Association*, 25(9):1161–1169, 2018b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

Wasswa William, Andrew Ware, Annabella Habinka Basaza-Ejiri, and Johnes Obungoloch. A review of image analysis and machine learning techniques for automated cervical cancer screening from pap-smear images. *Computer Methods and Programs in Biomedicine*, 165 (7):15–22, 2018.

World Health Organization. World Health Organization WHO, 2024. URL `https://www.who.int/`. Accessed: 2025-04-29.

Junyuan Xie, Ross B. Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International Conference on Machine Learning (ICML)*, pages 1070–1079. PMLR, 2016.

Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional GAN. *Advances in Neural Information Processing Systems*, 32, 2019.

Jianyu Zhao, Ying Zhang, Ying He, Xiaozhi Xie, and Jian Wu. Selective synthetic augmentation with HistoGAN for improved cervical histopathology image classification. *Medical Image Analysis*, 68:101919, 2021. doi: 10.1016/j.media.2020.101919.

Zhengyun Zhao, Qiao Jin, Fangyuan Chen, Tuorui Peng, and Sheng Yu. A large-scale dataset of patient summaries for retrieval-based clinical decision support systems. *Scientific Data*, 10(1):909, 2023.

Appendices

Appendix A

Preprocessing Resources and Vocabulary Specification

This appendix provides the complete resources required to reproduce the preprocessing pipeline.

Abbreviation Expansion Dictionary Clinical abbreviations were normalized using the following mapping:

Abbreviation	Expansion
w/	with
w/o	without
h/o	history of
s/p	status post
yo	years old
yr	year
yrs	years

Table A.1: Abbreviation Expansion Dictionary

Measurement Units Removed The following units were removed during preprocessing: bpm, cm, dl, g hr, kg, l, mcg meq, mg, ml, mm mmhg, mmol, μmol.

Domain-Specific Stop-word List The domain-specific stop-word list includes tokens identified as non-informative across clinical narratives. These consist of demographic, temporal, narrative, procedural, and generic clinical terms.

To illustrate the types of terms removed, a categorized summary of representative examples is provided below. The exhaustive, complete stop-word list is provided at the end of this section.

- **Demographic:** patient, male, female, age, years
- **Temporal:** day, month, week, time, later
- **Narrative verbs:** presented, showed, revealed, performed
- **Clinical workflow:** treatment, therapy, procedure, admission
- **Generic medical:** disease, pain, blood, test, symptoms
- **Modifiers:** normal, severe, mild, significant

Exclusion Criteria Summary Tokens were removed if they satisfied any of the following:

- Length < 4 characters
- Contained numeric characters
- Belonged to the measurement unit list
- Belonged to the stop-word set
- Appeared in >70% of documents (high-frequency)

- Appeared in <0.5% of documents (low-frequency)

The complete stop-word list The following is the complete domain-specific stop-word list used in preprocessing. These terms were removed as they were identified as non-informative across clinical narratives.

patient, year, years, old, male, female, age, weight, past, family, day, days, month, months, weeks, time, first, later, within, following, daily, one, two, three, showed, revealed, performed, presented, found, observed, reported, confirmed, started, treated, diagnosed, underwent, noted, hospital, admission, discharged, treatment, therapy, surgery, follow, admitted, department, emergency, referred, received, improved, done, made, continued, followed, initial, results, analysis, complete, total, disease, pain, blood, body, function, test, tests, signs, size, side, including, evidence, present, case, prior, week, per, min, rate, associated, second, low, small, large, could, given, history, medical, clinical, examination, physical, symptoms, diagnosis, findings, left, right, upper, lower, bilateral, normal, negative, positive, high, severe, mild, significant, increased, well, fig, figure, nthe, also, however, due, using, without, non, level, levels, range, count, demonstrated, developed, seen, evaluation, remained, improvement, course, since, used, included, based, presence, identified, consistent, general, along, care, previous, hours, four, post, placed, removed, woman, man, multiple, white, elevated, decreased, unremarkable, therefore, region, area, wall, type, detected, obtained, administered, presentation, taken, subsequently, respectively, approximately, stable, changes, procedure, clinic, condition, primary, local, period, free, anterior, posterior, distal, lateral, soft, acute, chronic, swelling, loss, bleeding, infection, fever, fluid, dose, oral, intravenous, postoperative, addition, although, previously, new, every, several, last, end, room, unit, discharge, score, response, abnormal, moderate, proximal, peripheral, back, areas, diameter, six, figures, syndrome, onset, flow, mg, ml, kg, g, mcg, cm, mm, dl, l, bpm, hr, x, meq, nshe, nhe, alt, fdg, igm, μmol, gram, use, next, good, like, according, despite, initiated, gradually, appeared, still, except, upon, may, single, returned, thus, common, show, point, would, able, taking, known, data, final, recent, related, third, around, considered, initially, control, status, minute, suggestive, report, result, change, defined, short, standard, whole, work, early, middle, light, decided, required, described, suggested, showing, rest, open, cause,

noticed, began, study, times, testing, less, throughout, closed, carried, main, material, order, life, team, state, process, number, earlier, completed, likely, successfully, obvious, round, term, effect, otherwise, near, limited, development, even, mean, cases, improve, came, almost, failed, different, added, decision, extension, direct, adjacent, minimal, phase, field, parameters, intervention, baseline, prominent, smooth, placement, panel, profile, guided, repair, invasive, criteria, mechanical, protocol, formation, underlying, component, advanced, combined, major, device, rehabilitation, investigation, induced, iii, drugs, boy, caused, section, complaint, health, significantly, monitoring, decrease, controlled, slight, recommended, clinically, human, height, differential, compared, portion, evaluated, extended, reveal, suffered, pre, thickness, quadrant, chain, base, amount, review, finally, doses, suggesting, perform, density, margin, compression, mid, involved, girl, university, elevation, image, resulted, worsened, rapid, sample, block, fixed, shaped, increasing, margins, sudden, eight, excised, sounds, delivery, technique, combination, visited, close, managed, finding, topical, origin, diseases, follows, raised, felt, seven, power, advised, basal, maximum, spontaneous, requiring, absent, sensory, saline, persisted, consciousness, midline, layer, differentiated, accompanied, affected, possibility, color, sensation, strength, opening, structures, mildly, disappeared, inguinal, normalized, suspicion, consisted, resulting, exhibited, gross, exposure, evident, superficial, scheduled, transverse, lead, entire, conducted, fine, changed, rapidly, available, additionally, illness, position, home, transferred, five, additional, denied, system, features, patients, possible, limits, twice, another, completely, reference, full, uneventful, shown, repeat, part, activity, secondary, ago, became, long, via, subsequent, indicated, clear, studies, internal, experienced, sided, risk, routine, measured, similar, usa, pattern, increase, investigations, immediately, operative, management, complaints, outpatient, recovery, administration, complained, specific, regular, minutes, resolved, exam, healthy, slightly, located, vital, resolution, intensive, occurred, hospitalization, active, marked, visit, prescribed, appearance, planned, poor, discontinued, parents, center, consent, absence, maintained, sent, values, value, view, achieved, support, assessment, duration, units, scale, length, generalized, recovered, informed, difficulty, intact, sign, surrounding, aspect, applied, died, effects, provided, led, indicating, consultation, institution, complex, containing, see, interval, mixed, access, adequate, starting, current, potential,

compatible, corrected, remaining, typical, excluded, tested, established, causing, visible, determined, collection, unable, regions, hard, thereafter, prevent, highly, composed, initiation, march, lack, class, activities, walking, stay, neither, degree, moreover, difficult, remarkable, functional, problems, immediate, node, air, systemic, complications, episodes, child, injury, growth, defect, pulse, tenderness, index, specimen, antigen, reduction, injection, space, deep, medication, medications, persistent, virus, infusion, motor, involving, extremities, mother, line, removal, irregular, hour, fraction, gait, frequency, flexion, screening, variant, ward, born

Appendix B

Clinical Keyword Dictionary

The complete clinical keyword dictionary used for specialty identification included 18 major medical specialties with 10–15 keywords each:

- **Radiology**: X-ray, CT scan, MRI, ultrasound, PET scan, fluoroscopy, contrast media, interventional, mammography, nuclear medicine, radiopaque, artifact, gantry, densitometry
- **Surgery**: excision, incision, repair, anesthesia, suture, scalpel, laparoscopy, biopsy, graft, intubation, debridement, resection, cautery, sterile field, preoperative
- **Cardiology**: cardiac, heart, myocardial, echocardiogram, ECG, angina, hypertension, arrhythmia, coronary, cardiomyopathy, stent, defibrillator, troponin, BNP, ejection fraction
- **Dermatology**: macule, papule, pustule, eczema, psoriasis, dermatitis, melanoma, lesion, epidermis, dermis, biopsy, erythema, pruritus, alopecia, keratin

- **Emergency Medicine**: triage, trauma, ACLS, CPR, intubation, shock, hemorrhage, laceration, syncope, overdose, anaphylaxis, sepsis, stabilization, GCS, STAT
- **Oncology**: cancer, tumor, malignant, chemotherapy, radiation, biopsy, metastasis, carcinoma, lymphoma, leukemia, oncology, adjuvant, palliative, remission, recurrence
- **Orthopedics**: fracture, bone, joint, spinal, arthritis, orthopedic, trauma, osteoporosis, skeletal, dislocation, cartilage, ligament, tendon, scoliosis, prosthesis
- **Gastroenterology**: abdominal, colon, endoscopy, hepatic, pancreatic, IBD, cirrhosis, hepatitis, colitis, gallbladder, GI, bowel, ulcer, reflux, diverticulitis
- **Pulmonology**: pulmonary, lung, respiratory, COPD, asthma, bronchitis, oxygen, pneumonia, tuberculosis, emphysema, ventilator, spirometry, hypoxia, pleural, bronchoscopy
- **Infectious Disease**: pathogen, antibiotic, antiviral, sepsis, zoonotic, resistance, incubation, asymptomatic, epidemic, vector, culture, titers, MRSA, fungal, transmission
- **Neurology**: neurological, brain, stroke, seizure, MRI brain, cognitive, dementia, migraine, neuropathy, Parkinson, Alzheimer, epilepsy, multiple sclerosis, tremors, aphasia
- **Endocrinology**: diabetes, thyroid, hormone, metabolic, insulin, glucose, endocrine, hypothyroid, hyperglycemia, diabetic, HbA1c, cortisol, pituitary, adrenal, gonad
- **Urology**: bladder, kidney, prostate, urinary, hematuria, cystoscopy, incontinence, nephrolithiasis, urethra, erectile dysfunction, catheter, ureter, lithotripsy, voiding, dysuria
- **Nephrology**: renal, dialysis, creatinine, GFR, electrolyte, urea, proteinuria, fistula, transplant, nephron, hypertension, uremia, hemodialysis, CKD, albumin

- **Ophthalmology**: eye, ocular, retinal, vision, cataract, glaucoma, ophthalmic, conjunctivitis, retinopathy, macular, cornea, iris, lens, vitreous, strabismus
- **OB/GYN**: pregnancy, fetus, labor, cesarean, ultrasound, menstruation, menopause, ovaries, uterus, cervix, prenatal, postpartum, ectopic, endometriosis, hysteroscopy
- **Pediatrics**: neonate, infant, adolescent, growth chart, immunization, congenital, developmental, pediatric, colic, croup, febrile, otitis media, ADHD, autism, bronchiolitis
- **Psychiatry**: mental health, anxiety, depression, bipolar, schizophrenia, therapy, psychosis, antidepressant, mood, cognitive, trauma, personality disorder, phobia, hallucination, mania

Appendix C

Record-level Classification by LLMs

Specialty	Patient ID	GPT-4o	GPT-5
Cardiovascular	62654	Onc-Path	Onc-Path
	60023	Correct	Correct
	139032	Correct	Correct
	84669	Correct	Renal-Meta
	42090	Correct	Correct
	68503	Onc-Path	Correct
	10088	Correct	Correct
	151873	Correct	Correct
	103638	Correct	Onc-Path
	69175	Correct	Correct
Derm-Immune	82803	Correct	Correct
	121260	Correct	Correct
	159318	Correct	Correct
	25030	Correct	Neuro-Pain
	165699	Correct	Correct
	159257	Onc-Path	Correct
	90306	Genet-Endoc	Correct
	1124	Correct	Rheu-Immune
	155724	Onc-Path	Correct
	148084	Genet-Endoc	Correct
Gastro	149095	Genet-Endoc	Genet-Endoc
	52145	Onc-Path	Infectious-Disease
	161890	Neurological	Neurological
	56214	Genet-Endoc	Onc-Path
	4541	Onc-Path	Neurological
	37640	Correct	Correct
	118160	Onc-Path	Musculoskeletal
	88882	Correct	Correct
	37705	Correct	Correct
	115278	Correct	Correct

Table C.1: Record-level Classification (Part 1)

Specialty	Patient ID	GPT-4o	GPT-5
Genet-Endoc	24467	Correct	Correct
	17841	Correct	Correct
	42006	Correct	Correct
	105877	Renal-Metabolic	Renal-Metabolic
	111090	Correct	Correct
	62432	Correct	Correct
	24945	Correct	Correct
	86865	Musculoskeletal	Correct
	60451	Correct	Correct
	79862	Correct	Correct
Inf-Disease	127353	Dermat-Immune	Derm-Onc
	23732	Correct	Correct
	141684	Correct	Correct
	127990	Gastro	Gastro
	103185	Gastro	Gastro
	149712	Correct	Correct
	121631	Correct	Correct
	55636	Correct	Correct
	110133	Correct	Correct
	90993	Correct	Correct
Genet-Endoc	24467	Correct	Correct
	17841	Correct	Correct
	42006	Correct	Correct
	105877	Renal-Metabolic	Renal-Metabolic
	111090	Correct	Correct
	62432	Correct	Correct
	24945	Correct	Correct
	86865	Musculoskeletal	Correct
	60451	Correct	Correct
	79862	Correct	Correct

Table C.2: Record-level Classification (Part 2)

Specialty	Patient ID	GPT-4o	GPT-5
Inf-Disease	127353	Dermat-Immune	Derm-Onc
	23732	Correct	Correct
	141684	Correct	Correct
	127990	Gastro	Gastro
	103185	Gastro	Gastro
	149712	Correct	Correct
	121631	Correct	Correct
	55636	Correct	Correct
	110133	Correct	Correct
	90993	Correct	Correct
Musculoskeletal	110805	Neurological	Genet-Endoc
	23887	Correct	Correct
	89655	Correct	Correct
	114951	Correct	Genet-Endoc
	133451	Correct	Correct
	165085	Derm-Immune	Derm-Immune
	112768	Derm-Immune	Derm-Immune
	35165	Neurological	Neurological
	133463	Correct	Correct
	137889	Correct	Correct
Neurological	72972	Correct	Correct
	42531	Correct	Correct
	8345	Correct	Genet-Endoc
	61730	Onc-Path	Onc-Path
	15629	Correct	Correct
	63297	Onc-Path	Correct
	142082	Onc-Path	Correct
	104683	Correct	Correct
	60846	Derm-Immune	Derm-Immune
	75131	Correct	Renal-Metabolic

Table C.3: Record-level Classification (Part 3)

Specialty	Patient ID	GPT-4o	GPT-5
Onc-Path	64555	Correct	Correct
	37206	Correct	Correct
	126668	Correct	Correct
	31073	Correct	Correct
	48515	Correct	Correct
	12604	Correct	Genet-Endoc
	54266	Correct	Correct
	131630	Correct	Gastro
	62745	Correct	Correct
	80030	Correct	Correct
Ophthalmology	165652	Correct	Correct
	159051	Correct	Correct
	98449	Correct	Correct
	102484	Correct	Correct
	133783	Correct	Correct
	121016	Correct	Correct
	136204	Correct	Correct
	71150	Correct	Correct
	26521	Correct	Correct
	121010	Correct	Correct
Renal-Metabolic	6185	Correct	Correct
	44961	Correct	Gastro
	26803	Correct	Correct
	164240	Genet-Endoc	Derm-Immune
	109984	Correct	Correct
	66975	Correct	Correct
	17684	Musculoskeletal	Musculoskeletal
	118814	Correct	Correct
	83888	Correct	Correct
	145177	Correct	Correct

Table C.4: Record-level Classification (Part 4)

Specialty	Patient ID	GPT-4o	GPT-5
Respiratory	89842	Correct	Correct
	51025	Correct	Correct
	132442	Correct	Correct
	54648	Cardiovascular	Cardiovascular
	78737	Neurological	Neurological
	3333	Correct	Correct
	51265	Correct	Renal-Metabolic
	155045	Genet-Endoc	Genet-Endoc
	116306	Correct	Correct
	48499	Correct	Correct

Table C.5: Record-level Classification (Part 5)

Appendix D

Computational Specifications and Optimizations

Hardware Specifications Experiments were conducted on a high-performance computing cluster with the following hardware configuration:

Component	Specification
CPU	Intel Xeon Gold 6248R (2.1 GHz base frequency)
RAM	128 GB DDR4 ECC (3200 MHz)
GPU	1 NVIDIA A100 80GB PCIe MIG 1g.10gb
CPU counts	128
GPU memory	10 GB
Storage	10 TB NVMe SSD RAID array
Network	100 GbE InfiniBand interconnect

Table D.1: Hardware Specifications for Experimental Evaluation

Software Environment The experimental software stack was carefully curated to ensure compatibility and reproducibility:

Component	Version/Specification
Operating System	Ubuntu 22.04 LTS
Python Distribution	Anaconda 2022.10 (Python 3.9.16)
Core Libraries	
torch	2.6.0 cu124
scikit-learn	1.2.0
sentence-transformers	2.2.2
transformers	4.27.4
pandas	1.5.3
numpy	1.24.2
scipy	1.10.0
Visualization	
matplotlib	3.7.0
seaborn	0.12.2
plotly	5.13.0
Development Tools	
JupyterLab	3.5.3
Git	2.34.1
Docker	20.10.21

Table D.2: Software Environment and Dependencies

Computational Optimizations Optimizations were implemented to improve computational efficiency:

Batch Processing Strategy BERT inference was optimized through batch processing with size 32, balancing GPU memory utilization (approximately 12 GB per batch) with computational throughput. This configuration achieved 95% GPU utilization while maintaining stable memory allocation across the $3,000$ document corpus.

Parallelization Implementation Feature extraction and similarity computations were parallelized using joblib with 32 worker processes, matching the available CPU core count.

This parallelization reduced computation time for pairwise similarity matrices from approximately 45 minutes to under 5 minutes.

Caching Architecture A hierarchical caching system stored intermediate results at three levels: (1) raw BERT embeddings in HDF5 format, (2) processed structured features in compressed NumPy arrays, and (3) similarity matrices in sparse matrix format. This caching eliminated approximately 70% of redundant computations across experimental runs.

Convergence Optimization Iterative algorithms employed adaptive convergence criteria with early stopping when improvement fell below 10^{-4} for three consecutive iterations. Maximum iteration limits of 300 ensured termination while maintaining solution quality, with actual convergence typically occurring within $50 - 100$ iterations.

Performance Benchmarks The following performance metrics were recorded during experimental execution:

Pipeline Stage	**Time (minutes)**	**Memory (GB)**
BERT Embedding Generation	18.5	14.2
Structured Feature Extraction	3.2	2.8
Feature Fusion	1.5	1.2
Clustering (K-means)	4.8	3.6
Evaluation Metrics	2.0	1.5
Total	**30.0**	**23.3**

Table D.3: Performance Benchmarks for Pipeline Components

These benchmarks represent average values across 10 experimental runs, with standard deviations below 5% of mean values, indicating consistent performance across executions.

Appendix E

Hyperparameter Settings

Complete hyperparameter settings for all experiments:

Parameter	Value
BERT Model	
Model name	BiomedNLP-PubMedBERT-base-uncased
Embedding dimension	768
Batch size	32
Pooling method	Mean pooling
Llama Model	
Model name	meta/Llama2
Embedding dimension	4096
Batch size	8
Pooling method	Max pooling
Structured Features	
TF-IDF max features	500
TF-IDF min document frequency	5
TF-IDF max document frequency	0.7
Specialty vector dimension	8
Feature Fusion	
Fusion weight (α)	0.3
PCA components	768
Normalization method	StandardScaler (z-score)
Clustering	
Algorithm	K-means
Cluster range evaluated	5-20
Optimal clusters	8
Initialization method	k-means
Number of initializations	10
Maximum iterations	300
Tolerance	1×10^{-4}
Evaluation	
Primary metric	Silhouette score
Secondary metrics	Davies-Bouldin index, Calinski-Harabasz index
Stability metric	Jaccard similarity
Statistical significance threshold	$p < 0.05$

Table E.1: Hyperparameter Settings for the Hybrid Clustering Pipeline

Appendix F

Evaluation Metrics Definitions

This appendix provides formal definitions for all evaluation metrics used throughout the thesis.

Segmentation Metrics For the evaluation of classification tasks, the following metrics were used:

- **Precision**: Measures the fraction of true positive cases out of all predicted as positive.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{F.1}$$

  where TP is the number of true positive and FP is the number of false positive.

- **Recall**: Measures the ability of a model to identify all relevant instances.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{F.2}$$

  where FN is the number of false negative records.

- **Accuracy**: Measures the fraction of correctly classified records out of all records.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{F.3}$$

  where TN is the number of true negative records.

- **Specificity**: Measures the proportion of actual negative instances that are correctly classified. This metric reflects the model's ability to avoid false alarms and is particularly relevant when distinguishing between closely related clinical conditions.

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{F.4}$$

- **F1 Score**: Harmonic mean of precision and recall, balancing both metrics.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{F.5}$$

Clustering Metrics For clustering tasks, the following metrics were used:

- **Silhouette Score**: Measures how similar an object is to its own cluster compared to other clusters (ranges -1 to $+1$).

- **Davies-Bouldin Index**: Measures the average similarity between each cluster and its most similar cluster (lower is better).

- **Calinski-Harabasz Index**: Ratio of between-cluster dispersion to within-cluster dispersion (higher is better).

- **Jaccard Similarity**: Measures stability of clustering across different initializations or data subsets.

- **Normalized Mutual Information (NMI)** Measures the agreement between the predicted clusters and reference labels.
- **Adjusted Rand Index (ARI)** Evaluates the similarity between two assignments in the same or different clusters in the predicted and reference labels.