MSc Thesis

Diagnosis of Diseases using AI Methods

Hybrid Multi-class Patient Classification on MIMIC-III with Bio_ClinicalBERT and MedGemma Evaluation

University of the Aegean

Internet of Things: Intelligent Environments in Next-Generation Networks

Petros Korlos

Supervisor: Assoc. Prof. Panagiotis Symeonidis

11 August 2025

Abstract

Electronic Health Records (EHRs) combine unstructured clinical narratives with structured measurements such as laboratory values and diagnostic codes. Leveraging both modalities is essential for building reliable clinical decision support tools. This thesis investigates multi-class patient classification into four mutually exclusive categories—*Healthy*, *Diabetic only*, *Cardiac only*, and *Both*—using data derived from the MIMIC-III database. We propose a hybrid model that fuses contextual text representations from **Bio_ClinicalBERT** with **numeric features** extracted via regular expressions from discharge summaries and structured laboratory signals (e.g., glucose, HbA1c, troponin, BNP). The fusion is performed by concatenating the pooled `[CLS]` embedding with the numeric vector and training a fully connected classification head with *weighted cross-entropy* to handle class imbalance.

The experimental protocol employs **5-fold cross-validation**. The hybrid model attains **Accuracy = 0.731**, **Macro-F1 = 0.730**, and **Macro-AUC = 0.890** on average across folds. Class-wise AUCs are consistently high (0.928, 0.892, 0.887, 0.853), with the comorbid *Both* class being the most challenging. Compared with a text-only baseline (Accuracy 0.656, Macro-F1 0.655, Macro-AUC 0.865), the inclusion of structured labs yields a **+7.5% relative gain in Macro-F1** and a **+2.5% increase in Macro-AUC**, confirming the benefit of multimodal integration.

Beyond the supervised experiments, we also conducted a focused evaluation of the MedGemma-27B medical language model on a binary cardiovascular disease detection task, using curated discharge summaries from the same cohort. In a small-scale, zero-shot setting (40 patients), MedGemma-27B achieved perfect discrimination, and attribution-based ablation together with feature-level inspection revealed that its decisions are strongly driven by explicit cardiovascular cues (e.g., hypertension, valvular surgery, cardiac function terms). In parallel, we deployed MedGemma as an LLM-based radiology assistant on a university server (25 GB VRAM) with a Gradio [1] demo interface, enabling interactive testing on X-ray/CT/MRI images. While MedGemma was not retrained within this work, these experiments demonstrate both the feasibility of incorporating multimodal foundation models into clinical workflows and their potential role as text-based diagnostic assistants. Overall, this thesis provides a reproducible pipeline for multi-class clinical classification and outlines concrete directions for calibration, temporal modeling, and multimodal fusion.

Executive Summary

Objective. We address the practical problem of assigning patients to one of four clinically meaningful categories—*Healthy*, *Diabetic only*, *Cardiac only*, *Both*—by integrating unstructured notes with structured laboratory information. The aim is to improve discrimination and clinical interpretability over text-only systems.

Data and Features. Using MIMIC-III, we derive labels from ICD-9 codes (`DIAGNOSES_ICD`) and extract text from discharge summaries (`NOTEEVENTS`). We design *regex-derived* binary and numeric indicators from the notes (e.g., diabetes mentions, maximum glucose/HbA1c) and aggregate *structured lab features* (e.g., glucose, troponin, BNP). These numeric signals complement contextual representations from Bio_ClinicalBERT.

Model. The architecture concatenates the `[CLS]` embedding (768-dim) with the numeric feature vector, feeding a fully connected head. We optimize with AdamW and *weighted cross-entropy*. Early stopping is applied per fold. Evaluation uses **5-fold cross-validation** with Accuracy, Macro-F1, and Macro-AUC.

Results. The hybrid model achieves **Accuracy 0.731**, **Macro-F1 0.730**, **Macro-AUC 0.890** (mean across folds). Class-wise AUCs exceed 0.85; the *Both* class remains hardest due to clinical overlap. Versus a text-only baseline (Accuracy 0.656, Macro-F1 0.655, Macro-AUC 0.865), structured labs provide a **+7.5%** relative Macro-F1 improvement and **+2.5%** Macro-AUC gain. In addition, a complementary evaluation of the MedGemma-27B medical language model on a binary cardiovascular classification task (40 curated admissions) yielded perfect discrimination (Accuracy, F1, and AUC all equal to 1.0 in this small test set). Subsequent ablation and feature-level analyses indicated that MedGemma's predictions are highly sensitive to explicit cardiovascular mentions, offering a qualitative interpretability layer on top of the supervised pipeline.

Deployment. Beyond the supervised classifier, we evaluated MedGemma-27B as a zero-shot cardiovascular text classifier and deployed MedGemma (LLM-only) on a university server with 25 GB VRAM and a Gradio demo interface for interactive radiology report generation. The text-based evaluation demonstrates that large medical foundation models can accurately detect cardiovascular disease directly from discharge summaries in a small, curated cohort, while the radiology prototype evidences a feasible path toward multimodal decision support that combines images and text.

Takeaways. (i) Multimodal fusion of text and labs is beneficial; (ii) high per-class AUCs indicate reliable ranking; (iii) future work should emphasize calibration, longer-context text encoders, temporal patient modeling, and principled hyperparameter search (e.g., nested CV); and (iv) foundation models such as MedGemma-27B can act as strong zero-shot baselines when carefully evaluated and interpreted, rather than used only as black-box demos. The provided code and figures enable reproducibility and extension.

Contents

Abstract i

Executive Summary ii

1 Introduction 1
- 1.1 Background . . . 1
- 1.2 Motivation . . . 1
- 1.3 Problem Definition . . . 2
- 1.4 Objectives . . . 2
- 1.5 Contributions . . . 3
- 1.6 Code and Data Availability . . . 3
- 1.7 Thesis Structure . . . 3

2 Related Work 5
- 2.1 Natural Language Processing in Healthcare . . . 5
- 2.2 BioBERT and ClinicalBERT . . . 5
- 2.3 Hybrid Models: Combining Text and Structured Data . . . 5
- 2.4 Multi-class Classification in Clinical NLP . . . 5
- 2.5 AI in Radiology and Multimodal Models . . . 6
- 2.6 Summary . . . 6

3 Data 7
- 3.1 MIMIC-III Database . . . 7
- 3.2 Tables Used . . . 7
- 3.3 Label Definition . . . 8
- 3.4 Class Balance . . . 8
- 3.5 Example Discharge Note . . . 8
- 3.6 Structured Laboratory Values . . . 11
- 3.7 Train/Validation/Test Split . . . 11
- 3.8 Cohort Construction . . . 12
  - 3.8.1 Main 4-class Cohort . . . 12
  - 3.8.2 40-patient Evaluation Subset for MedGemma . . . 12

4 Methods 13
- 4.1 Problem Formulation . . . 13
- 4.2 Data Sources . . . 13
- 4.3 Preprocessing . . . 13
- 4.4 Model Architecture . . . 14
- 4.5 Training Procedure . . . 15

4.6 Demonstration of MedGemma . . . 15
4.7 Evaluation with MedGemma . . . 15

5 The Proposed Method **17**
5.1 Feature-Level Interpretation of MedGemma-27B . . . 17

6 Experimental Setup **21**
6.1 Hardware and Environment . . . 21
6.2 Dataset Splitting and Class Balancing . . . 21
6.3 Training Settings . . . 22
6.4 Model Hyperparameters . . . 24
6.5 Evaluation Metrics . . . 24

7 Results and MedGemma Evaluation **26**
7.1 Overall Performance . . . 26
7.2 ROC-AUC by Class . . . 26
7.3 Classification Report . . . 26
7.4 Confusion Matrix . . . 27
7.5 ROC Curves . . . 28
7.6 Feature Correlation (Regex-derived) . . . 29
7.7 Analysis . . . 29
7.8 Classification Results – MedGemma-27B . . . 30
7.9 Comparative Evaluation on the 40-Patient Cohort . . . 31
7.10 Context-Grounded Ablation Study – MedGemma-27B . . . 32

8 Discussion **37**
8.1 Interpretation of Results . . . 37
8.2 Clinical Implications . . . 37
8.3 Strengths . . . 38
8.4 Limitations . . . 38
8.5 Ethical Considerations . . . 39

9 Future Work **40**
9.1 Data Quality and Cohort Expansion . . . 40
9.2 Feature Engineering and Calibration . . . 40
9.3 Modeling Improvements . . . 40
9.4 Evaluation Protocols and Hyperparameter Search . . . 41
9.5 Error Analysis and Interpretability . . . 41
9.6 Multimodal Extensions with MedGemma . . . 41
9.7 Operationalization, Ethics, and Governance . . . 42
9.8 Summary . . . 42

10 Conclusion **43**

A Appendix **45**
A.1 Regex Patterns for Feature Extraction . . . 45
A.2 Hyperparameters . . . 45
A.3 Hardware Specifications . . . 45
A.4 Sample Discharge Note Excerpt . . . 46

A.5 Executive Summary 46

References **48**

List of Figures

2.1 MedGemma Radiology Assistant – demonstration interface. 6

5.1 Sentence-level conceptual visualization. In the full four-sentence context (top), MedGemma-27B predicts cardiovascular disease, driven mainly by sentences describing an abdominal aortic aneurysm and a non-ST elevation myocardial infarction (S1, S2). After removing these key sentences (bottom), only respiratory and rehabilitation information (S3, S4) remains, and the model's output shifts to "No cardiovascular disease" / "Insufficient information". 19

7.1 Confusion matrix of the four-class classification (example from one validation fold). 27
7.2 One-vs-rest ROC curves for the four classes (mean across folds). 28
7.3 Correlation heatmap among regex-derived features extracted from text. 29
7.4 Confusion Matrix – MedGemma-27B (Binary Cardiovascular Classification). . 31
7.5 Unified ROC curves for cardiovascular disease classification on the 40-patient cohort. The Hybrid Bio_ClinicalBERT model is shown in blue, while MedGemma-27B is shown in orange. Confusion Matrix only from Hybrid Bio_ClinicalBERT 32
7.6 Confusion matrices for MedGemma-27B after removing the top-k most influential sentences from the context ($k = 2, 4, 6$). Each matrix is computed on the same 40-patient cohort (20 with cardiovascular disease and 20 without). As k increases, the number of false negatives grows while the number of false positives remains zero, indicating a progressive loss of sensitivity but preserved specificity. 34

List of Tables

3.1 Cohort characteristics for the 40-patient evaluation set. 12

4.1 Models and training/inference configurations. 15

5.1 Representative linguistic features influencing MedGemma-27B predictions. . . 18

6.1 Hyperparameters of the Hybrid model (Bio_ClinicalBERT + MLP numeric head). 23
6.2 Configuration of the MedGemma-27B zero-shot pipeline (ablation-ready). . . . 23

7.1 Overall performance of the hybrid model (5-fold cross-validation). 26
7.2 Class-wise ROC-AUC scores (mean across 5 folds). 26
7.3 Classification report (mean across 5 folds). 27
7.4 Performance metrics for MedGemma-27B on 40 patient samples. 30
7.5 Hybrid vs. MedGemma-27B on the same 40-patient cohort 32
7.6 Average top-k log-probability drop and flip rate for MedGemma-27B (20 patients). 33

A.1 Training hyperparameters. 45

Chapter 1

Introduction

1.1 Background

The rapid expansion of digital healthcare has led to a massive growth in electronic health records (EHRs). These records contain both structured data, such as laboratory values and diagnostic codes, and unstructured data, such as free-text clinical notes and discharge summaries. Extracting meaningful insights from this data has become a cornerstone of modern clinical research and decision support. Artificial Intelligence (AI), and more specifically Natural Language Processing (NLP), has demonstrated remarkable capabilities in automating tasks such as patient classification, outcome prediction, and clinical decision support.

Large Language Models (LLMs) like BERT and its medical variants (e.g., BioBERT, ClinicalBERT) have enabled significant progress in understanding clinical text. Meanwhile, the availability of open datasets such as MIMIC-III has created new opportunities for building reproducible and clinically relevant machine learning systems. In parallel, multimodal systems that combine text, numerical features, and imaging have gained traction, improving disease diagnosis and interpretability.

1.2 Motivation

Cardiovascular disease and diabetes are two of the most prevalent chronic diseases worldwide. They frequently co-exist in patients, leading to increased morbidity, higher hospitalization rates, and more complex care pathways. Early identification of patients with these conditions, using automated systems that process both textual and structured hospital data, can help healthcare providers improve patient care and resource allocation.

However, clinical notes are long, noisy, and heterogeneous. They often contain abbreviations, misspellings, and fragmented information. Traditional machine learning approaches cannot efficiently capture their complexity. Recent progress in transformer-based models, particularly those pre-trained on biomedical corpora, provides a promising path for addressing this challenge. Combining such models with structured features derived from laboratory results and diagnosis codes enables a more holistic representation of patient health.

Beyond classical text-based methods, the emergence of large, multimodal medical models such as MedGemma introduces new opportunities for medical text understanding, context reasoning, and visual–language integration. Evaluating such models on real hospital discharge summaries, and probing their decision behaviour through ablation and feature-level analyses, forms an important part of this thesis.

1.3 Problem Definition

This thesis addresses the task of multi-class patient classification. Specifically, we classify hospital admissions into four mutually exclusive categories:

1. Healthy: neither diabetes nor cardiovascular disease,
2. Diabetic only,
3. Cardiac only,
4. Both conditions.

The problem is challenging due to:

- The variability and length of discharge summaries,
- The overlap between conditions (e.g., diabetes and cardiovascular disease often co-occur),
- Class imbalance in raw data distributions,
- The need to combine heterogeneous sources: free-text notes and structured laboratory results.

1.4 Objectives

The goals of this thesis are:

- To preprocess MIMIC-III data and build a balanced, patient-level dataset for multi-class classification.
- To design a hybrid model that integrates Bio_ClinicalBERT embeddings with numeric features extracted via regex and structured laboratory values.
- To train and evaluate the system on four classes, reporting Accuracy, Macro-F1, and ROC-AUC metrics.
- To analyze model errors and identify challenges in distinguishing between overlapping disease classes.
- To investigate the role of large medical language models by:
  - evaluating **MedGemma-27B** as a zero-shot text classifier for cardiovascular disease, and
  - deploying MedGemma in a radiology demo for MRI, CT, and X-ray reporting.

1.5 Contributions

The contributions of this work can be summarized as follows:

1. Construction of a reproducible multi-class classification pipeline on MIMIC-III.
2. Integration of textual embeddings and structured features in a unified architecture.
3. Empirical evidence that structured laboratory features significantly improve classification performance over text-only baselines.
4. Quantitative evaluation of **MedGemma-27B** as a zero-shot medical text classifier for cardiovascular disease, using curated discharge summaries from MIMIC-III.
5. Development of a demo system using MedGemma for radiology report generation, illustrating the potential of AI-assisted diagnosis in multimodal healthcare scenarios.

1.6 Code and Data Availability

To support reproducibility and facilitate further research, the code and processed datasets used in this thesis are publicly available.

- **Source Code**: The full implementation of the hybrid Bio_ClinicalBERT model, preprocessing pipeline, evaluation scripts, and visualization utilities is available at: `https://github.com/USERNAME/REPOSITORY`
- **Dataset 1 (4-class cohort)**: The processed and balanced four-class admission-level dataset derived from MIMIC-III (excluding raw clinical notes) is available at: `https://physionet.org/content/mimiciii/1.4/`
- **Dataset 2 (40-patient evaluation subset)**: The discharge summaries are encoded into fixed-dimensional vector representations using Bio_ClinicalBERT, and these embeddings are used as inputs to the multiclass classification model. Raw clinical text is not released. The curated 40-admission cohort used for the MedGemma-27B - Hybrid Bioclinical cardiovascular disease evaluation is available at: `https://zenodo.org/record/YYYYYYY`. This archive also includes ablation-level metadata and evaluation outputs for the 20 cardiovascular disease–positive patients used in the context-grounded ablation study, without releasing raw clinical text.

1.7 Thesis Structure

The remainder of this thesis is organized as follows:

- **Chapter 2** reviews related work in clinical natural language processing, hybrid models that combine text with structured data, multimodal foundation models, and AI applications in healthcare.
- **Chapter 3** describes the dataset, cohort construction, and preprocessing steps. It details how discharge summaries and structured laboratory features are extracted from MIMIC-III and introduces the 40-patient evaluation subset.

- **Chapter 4** presents the methods used in this thesis, including the hybrid classifier that integrates Bio_ClinicalBERT text embeddings with laboratory features and the zero-shot cardiovascular disease classifier based on MedGemma-27B.
- **Chapter 5** introduces the proposed interpretability perspective, termed *Counterfactual Explanations for Justifying Automated Differential Diagnosis of Diseases*. It focuses on feature- and sentence-level analyses of MedGemma-27B to identify key words and sentences that support specific diagnostic decisions.
- **Chapter 6** outlines the experimental setup, including dataset splitting, class balancing, training procedures, hyperparameter configurations, and evaluation metrics for both the hybrid classifier and MedGemma-27B.
- **Chapter 7** reports and analyzes the results of the multi-class classification experiments and the binary cardiovascular disease task. It provides a dedicated MedGemma evaluation, including attribution-based ablations, feature-level interpretations, and a comparative analysis on the 40-patient cohort.
- **Chapter 8** discusses the main findings, strengths, and limitations of the proposed approach, as well as ethical and practical considerations for deploying such models in clinical settings.
- **Chapter 9** outlines future work directions, with an emphasis on extending the current framework to multimodal differential diagnosis by incorporating imaging data via MedGemma-based models and improving explanation mechanisms.
- **Chapter 10** concludes the thesis by summarizing the contributions and reflecting on the broader implications of the results.

Chapter 2

Related Work

2.1 Natural Language Processing in Healthcare

The use of NLP in healthcare has advanced rapidly in recent years. Traditional approaches relied on bag-of-words models or simple embeddings, which were insufficient for capturing the complex semantics of clinical text. The introduction of transformer-based architectures, particularly BERT, revolutionized text representation. Variants such as BioBERT and ClinicalBERT were pre-trained on biomedical and clinical corpora, enabling improved performance on tasks such as named entity recognition, relation extraction, and clinical classification.

2.2 BioBERT and ClinicalBERT

BioBERT [2] was trained on large-scale biomedical corpora including PubMed abstracts and PMC full-text articles, achieving state-of-the-art results on biomedical NLP tasks. ClinicalBERT [3] further adapted BERT to the clinical domain using the MIMIC-III notes, which closely resemble the data used in this thesis. These models demonstrate the importance of domain adaptation in achieving reliable performance in clinical applications.

2.3 Hybrid Models: Combining Text and Structured Data

While language models excel at text understanding, many clinical prediction tasks benefit from structured features such as laboratory measurements, demographics, and vital signs. Prior work has shown that combining text embeddings with structured data improves classification and prediction tasks [4], [5]. Hybrid models provide richer patient representations and often outperform text-only or structured-only baselines.

2.4 Multi-class Classification in Clinical NLP

Most clinical NLP work has focused on binary classification tasks, such as identifying whether a patient has diabetes or not. Multi-class classification is more complex because classes may overlap semantically or clinically. For example, distinguishing between “diabetic only” and “both diabetic and cardiac” requires nuanced understanding. Previous studies on ICD coding [6] and phenotype classification [7] highlight the difficulty of multi-class classification in EHRs and motivate the hybrid approach used in this thesis.

2.5 AI in Radiology and Multimodal Models

Beyond text and structured data, AI has made significant progress in medical imaging. CheXNet [8] demonstrated that convolutional neural networks can outperform radiologists in pneumonia detection on chest X-rays. More recently, multimodal foundation models such as MedGemma have been introduced, which extend large language models to handle text and imaging jointly.

MedGemma combines a vision encoder, which extracts feature representations from radiological images, with a large language model (Gemma) that processes textual prompts. This architecture enables tasks such as report generation, abnormality detection, and diagnostic reasoning by aligning imaging and textual modalities. It exemplifies the broader trend of moving from unimodal to multimodal diagnostic assistants.

In the context of this thesis, MedGemma was not retrained but deployed as a demonstration prototype. The model runs on a university server equipped with **25 GB VRAM** and is exposed through a lightweight **Gradio interface**. This setup allows testing with MRI, CT, and X-ray scans, producing structured diagnostic summaries in real time. While the main focus of this thesis is the supervised multiclass classification model trained on MIMIC-III data, the MedGemma deployment illustrates the potential for integrating large multimodal models into clinical workflows.

In parallel with imaging applications, medical foundation models are increasingly used as text-only clinical assistants, answering diagnostic questions directly from discharge summaries and progress notes. This raises questions about how such models use clinical evidence and how sensitive their predictions are to the presence or absence of specific phrases. In this thesis, **MedGemma-27B** is therefore also evaluated as a zero-shot cardiovascular disease classifier on curated discharge summaries, and its behaviour is further analyzed using attribution-inspired ablation and feature-level inspection (see Chapter 7, especially Section 7.8).

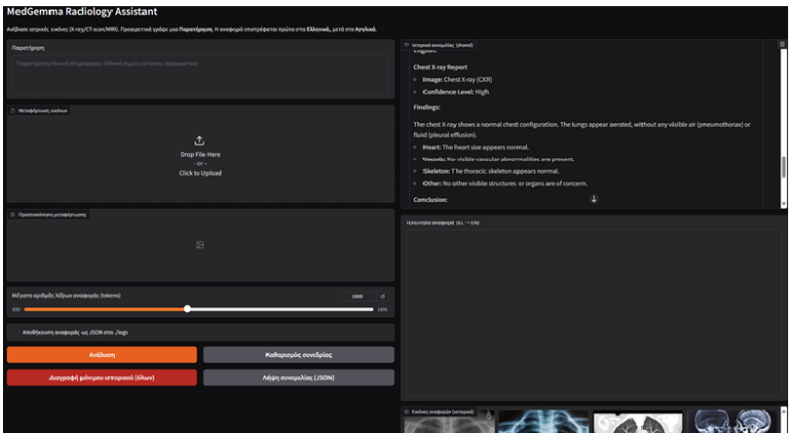


Figure 2.1: MedGemma Radiology Assistant – demonstration interface.

2.6 Summary

This thesis builds on prior work in clinical NLP and multimodal learning. By leveraging Bio_ClinicalBERT for text, structured laboratory values for numeric signals, and MedGemma for radiology, it aims to explore the practical integration of multiple modalities into patient classification and diagnostic support. The inclusion of a deployed MedGemma prototype highlights the feasibility of bringing multimodal AI research closer to clinical usability.

Chapter 3

Data

3.1 MIMIC-III Database

The Medical Information Mart for Intensive Care (MIMIC-III) [9] is a large, freely accessible database comprising de-identified health-related data associated with over 40,000 patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. It contains both structured and unstructured data, including:

- Demographics
- Laboratory measurements
- Vital signs
- Diagnosis codes (ICD-9)
- Free-text notes such as discharge summaries

This richness makes MIMIC-III an excellent benchmark for clinical NLP and multimodal learning research.

3.2 Tables Used

For this thesis, we use the following tables:

- **DIAGNOSES_ICD:** Contains ICD-9 diagnosis codes associated with hospital admissions.
- **NOTEEVENTS:** Contains free-text notes, including discharge summaries.
- **D_LABITEMS:** Defines laboratory test identifiers.
- **LABEVENTS:** Contains laboratory results for each hospital admission.

3.3 Label Definition

Two conditions are targeted:

- **Diabetes:** Any ICD-9 code starting with 250 (250.xx).
- **Cardiac disease:** ICD-9 codes 410–414 (ischemic heart disease) or 428 (heart failure).

From these, four mutually exclusive categories are defined:

1. 0 – Healthy: neither diabetes nor cardiac disease,
2. 1 – Diabetic only,
3. 2 – Cardiac only,
4. 3 – Both conditions.

3.4 Class Balance

Because disease prevalence is uneven in raw MIMIC data, the cohort is balanced by downsampling the majority classes to match the smallest class. The final dataset includes approximately 12,032 admissions (around 3,000 per class).

3.5 Example Discharge Note

Discharge summaries in MIMIC-III are semi-structured documents authored by clinicians at the end of an admission. While their layout varies across services and time, they typically contain (some subset of) the following sections: *Header/Meta*, *Chief Complaint*, *Major Procedures*, *History of Present Illness (HPI)*, *Past Medical/Surgical History (PMH/PSH)*, *Social History*, *Physical Examination*, *Pertinent Results* (labs, imaging, ECG/ABG), *Brief Hospital Course*, *Medications on Admission*, *Discharge Medications*, *Disposition & Facility*, *Discharge Diagnoses*, *Condition/Instructions/Follow-up*. The example below illustrates a complete note; square-bracketed tokens (e.g., `[** ... **]`) denote de-identified content.

Full example note (de-identified).

```
subject id  5434  HADM ID 117196
2103-05-11  Admission Date:  [**2103-5-4**]  Discharge Date:  [**2103-5-11**]

Service: MEDICINE

Allergies: Penicillins

Attending:[**First Name3 (LF) 898**] Chief Complaint: diarrhea

Major Surgical or Invasive Procedure: Brief Intubation Thoracentesis

History of Present Illness: 83 F called EMS because she felt "not good
for too long." Pt denies weakness. She states that she last felt well
until about 4 weeks ago when she returned from a visit to her daughter
in [**Name (NI) 5256**]. She states that she just felt tired after getting
back. Of note, 1 month ago she noted black stools x 2 weeks that went
away on its own. She did not report this to her doctor. Several weeks
ago she began to have diarrhea off and on. Dark brown and voluminous.
```

1-2 weeks ago she developed a cough, occasionally productive of yellow sputum. She has also had some shortness of breath. Today the diarrhea returned and she had 2 large B<s and then called EMS. She denies F/C, N/V, CP. She states she has had a normal appetite. Denies rash. No recent antibiotic use. No abd or back pain.

When the EMS arrived, they found her VS BP 152/82 HR 102, irregular, RR 20 93and transported to ED. In the ED, she had a CXR which suggested LLL pna and was started on Levofloxacin. She then developed respiratory distress and was intubated. A CT chest was negative for PE but notable for ....

ROS also notable for progressive Lower extremity edema. + erythema of lower extremity. Pt feels her legs are a "weak part of her body."

Past Medical History: PMH: PCP [**Name9 (PRE) **] [**Name Initial (PRE) **] Fib dx'd [**2100**], not on coumadin Right corona radiata stroke [**2096**], with persistent left upper arm rigidity Hypothyroidism s/p thyroidectomy RA OA RLL PNA [**2100**]/ LLL PNA [**2100**] requiring hospitalization Hyperlipidemia [**2100**] echocardiogram EF 55

PSH: s/p b/l knee replacement

Social History: Social Hx: Vet. h/o smoking, lives with daughter.

Physical Exam: (Initially by medicine team in ED earlier in day): -VS 98.4 HR 98 BP 131/94 RR 24 O2 95-Gen: elderly female resting in bed NAD -HEENT: PERL. EOMI. + lentis senilis. MM dry. -Neck: JVP flat. horizantal well-healed scar immediatly superior to clavicle. -CV: irregulary irregular. no m/r/g. -Lungs: + dullness to percussion. absent breath sounds LLL. no egophany. -Abd: active BS. soft NT. ND. no masses. -Extr: 2+ pitting, erythematous edema to knee. TTP. PT 1+ B/L. -Neuro: CN II-XII grossly intact. slight biceps weakness of left upper extr. toes downgoing b/l. sensation intact diffusely. -Rectal: normal tone, guaiac negative

(upon MICU eval): T 99 BP 96/50 P87 Sat 100-Gen: sedated, intubated -HEENT: PERRL, EOMI -Neck: JVP -CV: irregularly irregular, no m/r/g -Pulm: CTA anteriorly although decreased BS at bases -Abd: s/nt/no grimacing, nd, +BS -Ext: 2+ pitting edema bilateral lower ext, erythematous

Pertinent Results: (at admission) -WBC 9 w/77N, 14L, 6M, 1.8E -[**Year (4 digits) **] 39.9, Plt 389 -Na 138, K 4.6, Cl 103, bicarb 24, BUN 15, creat 0.8, glc 110, Ca 9, Mg 2, Phos 2.9 -TSH 3.0 -BNP 1372 -INR 1.2 -lactate 1.6->1.2 -UA: contaminated -EKG initially: irregularly irregular 120s, occ PVCs nl axis, good R-wave progression. normal intervals. No ST segment changes. -EKG ?pre-intubation: irregularly irregular 120s, occ PVCs, very wavy baseline, no acute change. -CXR (pre-intubation): blunting of left costophrenic angle on lateral ? effusion vs consolidation. Otherwise overall clear with no upper zone redistribution -CXR (post-intubation): increased upper zone redistribution, hilar fullness and interstitial edema c/w CHF -ABG post-intubation 7.40/43/153 -Chest CTA (prelim read): no PE, patchy ground glass opacities with bilat pleural effusions and +consolidation c/w CHF and multifocal PNA

Brief Hospital Course: 1. Pneumonia: LLL consolidation apparent on both AP and lateral views on admission. ... 2. Atrial Fibrillation: ... risks of not being on this were explained ... 3. CHF: ... diuresed ... echo EF 45%, 1+ MR, 1+ TR, mild PA sys HTN ... thoracentesis ... 4. Black

```
stool: ... outpatient colonoscopy ... 5. Hypothyroidism: ... TSH normal
... 6. Aspiration risk/S&S: ... video swallow outpatient ...

Medications on Admission: Synthroid 0.75, Lasix (past), Aldactone (past),
Fosinopril (past)

Discharge Medications: 1. Acetaminophen 325 mg ... 2. Levothyroxine 75
mcg ... 3. Aspirin 325 mg EC ... 4. Heparin 5000U TID ... 5. Senna 8.6
mg BID ... 6. Docusate 100 mg BID ... 7. Pantoprazole 40 mg Q24H ...
8. Ipratropium nebs Q6H PRN ... 9. Diltiazem 90 mg QID ... 10. Ceftriaxone
1 g IV Q24H 11. Azithromycin 500 mg OD x3d 12. Furosemide 40 mg IV OD
13. Outpatient Speech/Swallow Therapy

Discharge Disposition: Extended Care

Facility: [**Hospital3 **] Center/MACU

Discharge Diagnosis: Primary: Multifocal Pneumonia, CHF Secondary: AFib,
hyperlipidemia, h/o CVA, hypothyroidism, RA, OA

Discharge Condition: Stable

Discharge Instructions: ... discuss anticoagulation for AFib with PCP
...

Followup Instructions: Follow-up with PCP [**Name Initial (PRE) 176**]
within 1 week; consider Coumadin; consider ACE inhibitor.
```

How we use each section in the pipeline. Below we summarize typical content and downstream use in our NLP pipeline (exact fields vary by note).

- **Header/Meta** (subject/HADM, dates, service, allergies, attending). Used for indexing admissions, computing length-of-stay, and excluding allergy-confounded medications.
- **Chief Complaint.** Short free-text cue for initial symptom framing (seed features for triage-like signals).
- **Major Procedures.** High-salience interventions (e.g., intubation, thoracentesis) that often correlate with disease severity and ICU admission.
- **History of Present Illness (HPI).** Main narrative of symptom onset, timeline, triggers, and initial ED course; richest source of disease-specific lexical evidence.
- **PMH/PSH.** Chronic comorbidities and prior operations (e.g., AFib, hypothyroidism) that strongly inform Cardiac/Diabetic labels and longitudinal risk.
- **Social History.** Contextual risk factors (e.g., smoking) that support differential diagnosis but are often weakly weighted.
- **Physical Examination.** Objective bedside findings; helpful for disambiguating respiratory vs. cardiac etiologies (e.g., edema, JVP, auscultation).
- **Pertinent Results.** Structured or semi-structured evidence (labs, ECG/ABG, imaging) that anchors clinical hypotheses (e.g., BNP, EF, pleural effusions).
- **Brief Hospital Course.** Day-by-day evolution, treatments, and response; high-yield for consolidating final diagnoses (e.g., CHF, AFib) and outcomes.

- **Medications on Admission / Discharge Medications.** Signals of chronic disease (e.g., levothyroxine, rate control for AFib) and therapeutic intent at discharge.
- **Disposition/Facility & Diagnoses/Condition/Follow-up.** Ground-truth style labels and outcomes; used for supervision and error analysis.

De-identification notes. Tokens in `[** ... **]` indicate PHI removed by the dataset provider; we treat them as neutral placeholders and exclude them from any entity-level attribution.

Caveat on variability. The exact section titles, ordering, and granularity vary across admissions and services; missing sections are common. Our preprocessing therefore uses robust section splitters and fallbacks (e.g., merging "Hospital Course" and "Assessment/Plan" if only one is present), ensuring consistent downstream feature extraction.

3.6 Structured Laboratory Values

From `LABEVENTS`, we select key biomarkers relevant to diabetes and cardiovascular disease:

- Glucose
- HbA1c
- Creatinine
- Estimated GFR
- Troponin
- BNP

For each biomarker, the maximum value per admission is extracted. Missing values are imputed as zero.

3.7 Train/Validation/Test Split

Splits are made at the patient level to avoid information leakage. The dataset is divided into:

- 60% training,
- 20% validation,
- 20% test.

This ensures that notes from the same patient do not appear in multiple splits.

3.8 Cohort Construction

3.8.1 Main 4-class Cohort

The main cohort is constructed at the hospital admission level (`HADM_ID`), linked to patients via `SUBJECT_ID`. For each admission, we extract the final discharge summary (the last note according to `CHARTDATE`) and the corresponding ICD-9 diagnosis codes from `DIAGNOSES_ICD`. Using the label definition from the previous section, each admission is assigned to one of the four mutually exclusive classes: Healthy, Diabetic only, Cardiac only, or Both.

Because the raw class distribution in MIMIC-III is highly imbalanced, we downsample the majority classes so that all four classes contain approximately the same number of admissions. After downsampling, the final 4-class dataset consists of about 12,032 admissions (roughly 3,000 per class). This balanced cohort is used for training and evaluating the supervised hybrid model. The train/validation/test split described in the previous section is applied at the patient level (i.e., `SUBJECT_ID`) to avoid information leakage across splits.

3.8.2 40-patient Evaluation Subset for MedGemma

In addition to the main 4-class cohort, we construct a smaller, manually curated evaluation subset aimed at analysing MedGemma-27B and comparing it to the hybrid model on a common set of cases. This subset consists of 40 unique patients (40 admissions), evenly split between cardiovascular and non-cardiovascular cases.

More concretely, we select admissions from the extremes of the 4-class label space: 20 admissions labelled as *Healthy* and 20 admissions labelled as *Cardiac only*. These are then relabelled into a binary outcome: No-CVD vs. CVD. This design yields a balanced and clinically interpretable binary cohort for the cardiovascular experiments.

Table 3.1: Cohort characteristics for the 40-patient evaluation set.

Characteristic	**Value**
Overview	
Patients (unique)	40
Admissions	40
Class distribution (4-class labels)	
Healthy	20 (50.0%)
Cardiac only	20 (50.0%)
Binary label (CVD vs No-CVD)	
No-CVD (`cardio`=FALSE)	20 (50.0%)
CVD (`cardio`=TRUE)	20 (50.0%)
Demographics	
Sex: Female	15 (37.5%)
Sex: Male	25 (62.5%)
Age (years)	Mean ≈ 55, Min $= 1$, Max $= 90$
Notes	
Source	Manually curated list of 40 admissions (20 No-CVD, 20 CVD) from MIMIC-III discharge summaries and services (`service` field).

Chapter 4

Methods

4.1 Problem Formulation

The task addressed in this thesis is a four-class patient classification problem using electronic health record (EHR) data. The classes are:

1. **Healthy**: neither diabetes nor cardiac disease.
2. **Diabetic only**: patients with diabetes but no cardiac disease.
3. **Cardiac only**: patients with cardiac disease but no diabetes.
4. **Both**: patients with comorbid diabetes and cardiac disease.

This setup extends beyond binary classification and requires models capable of capturing nuanced distinctions between overlapping disease categories.

4.2 Data Sources

The data were derived from the publicly available MIMIC-III database, which contains structured laboratory measurements, diagnostic codes (ICD-9), and unstructured clinical notes. Specifically:

- **DIAGNOSES_ICD** provided ICD-9 codes used to assign class labels.
- **NOTEEVENTS** supplied discharge summaries used for text-based features.
- Structured laboratory values were extracted from relevant EHR tables (glucose, HbA1c, creatinine, GFR, troponin, BNP).

4.3 Preprocessing

Preprocessing involved several steps:

- Regex-based extraction of relevant sections from discharge summaries (Past Medical History, Labs, Medications).
- Normalization of laboratory values and removal of extreme outliers.

- Construction of binary flags and maximum values to capture clinically relevant signals. The final set of features includes:
  - `has_ins`: indicator for presence of insulin or antidiabetic treatment mentions.
  - `max_glc`: maximum glucose value observed.
  - `max_hba`: maximum HbA1c value observed.
  - `has_creat`: indicator for presence of creatinine mentions.
  - `has_gfr`: indicator for presence of GFR mentions.
  - `max_creat`: maximum creatinine level observed.
  - `max_gfr`: maximum GFR value observed.
  - `has_tro`: indicator for presence of troponin mentions.
  - `has_bnp`: indicator for presence of BNP mentions.
  - `max_tro`: maximum troponin value observed.
  - `max_bnp`: maximum BNP value observed.

These features were combined with text-derived embeddings to form the multimodal patient representation.

4.4 Model Architecture

The classification model integrates text and structured data:

- A **Bio_ClinicalBERT** encoder produces contextualized embeddings from discharge summaries.
- Numeric features (listed above) are concatenated with the pooled `[CLS]` embedding.
- A fully connected neural head processes the combined vector into four output logits.
- A weighted cross-entropy loss addresses class imbalance.

In addition to this supervised model, the present work also explores large multimodal medical models such as **MedGemma-27B**, which can reason jointly over textual and visual medical evidence. While Bio_ClinicalBERT forms the main classification backbone, MedGemma is used later for qualitative evaluation and interpretability studies (see Chapters 7 and 5).

Table 4.1: Models and training/inference configurations.

Model	Supervision / Task	Key Settings	Hardware
Hybrid (Bio_ClinicalBERT + MLP)	Supervised / 4-class	Tokenizer: emilyalsentzer/Bio ClinicalBERT max_len=512 AdamW (lr=2e-5), batch=4 epochs=2 early stopping numeric features: regex + structured (max per admission)	GPU: NVIDIA RTX 4070 (12 GB VRAM); CPU: Intel i5-8600K; RAM: 16 GB
MedGemma-27B (Text-IT)	Zero-shot / Binary (CVD vs No-CVD)	Prompted yes/no greedy decoding (temperature=0.0, top-p=1.0), max_new_tokens=256 attribution-based ablation (k=2, 4, 6) for analysis	GPU: NVIDIA RTX (24 GB VRAM), university server

4.5 Training Procedure

To keep the experimental design focused and reproducible, we report two model families: (i) a supervised hybrid classifier that integrates Bio_ClinicalBERT text embeddings with structured laboratory features for the four-class task, and (ii) a zero-shot MedGemma-27B text-only setup for binary cardiovascular disease (CVD) detection. Earlier references to a "BioClinicalBERT + Logistic Regression" baseline were removed, as the final pipeline trains a small feed-forward head (MLP) on top of the encoder and numeric features.

4.6 Demonstration of MedGemma

In addition to the supervised classification experiments, a prototype deployment of **MedGemma** was carried out to illustrate multimodal diagnostic support. MedGemma was not retrained within this work, but was instead installed on a university server equipped with **25 GB VRAM**. A simple **Gradio** web interface was implemented, enabling test uploads of radiological scans (MRI, CT, X-ray) and generation of structured diagnostic outputs.

This demonstration serves as a parallel proof-of-concept for integrating large multimodal models into clinical workflows. The evaluation of MedGemma's reasoning, including attribution-based ablations and feature-level analyses, is presented in Chapters 5 and 7.

4.7 Evaluation with MedGemma

Beyond the supervised setup, MedGemma was also utilized to assess contextual reasoning on cardiovascular diagnostic narratives. Specifically, we performed:

- **Classification Evaluation:** Comparison of MedGemma-27B predictions on patient-level narratives.
- **Ablation Study:** Stepwise removal of the most influential sentences to observe prediction sensitivity.
- **Feature-Level Interpretation:** Qualitative analysis of which phrases drive decision changes.

This extended evaluation bridges traditional supervised modeling with modern multimodal large-language models, demonstrating how interpretability methods can reveal the linguistic and contextual signals guiding clinical reasoning.

Chapter 5

The Proposed Method

5.1 Feature-Level Interpretation of MedGemma-27B

In the previous chapters, we introduced the cohort construction, feature extraction pipeline, and the supervised and zero-shot models used in this thesis. In particular, we described a hybrid classifier that combines Bio_ClinicalBERT text embeddings with structured laboratory features, and a zero-shot cardiovascular disease classifier based on MedGemma-27B. While these models provide strong predictive performance, their usefulness in clinical practice ultimately depends on our ability to explain *why* a given diagnosis was selected.

In this chapter, we present our interpretability perspective, which we term *Counterfactual Explanations for Justifying Automated Differential Diagnosis of Diseases*. The central idea is that, given a patient's symptoms, clinical findings, and hospital course in free text form, the model must select the most likely disease among several plausible alternatives and justify this choice by identifying the *key words and sentences* that support it. Conversely, we are interested in how the prediction would change if these key phrases were removed or altered, thereby revealing why competing diagnoses with similar symptom profiles are ultimately ruled out.

The remainder of this chapter focuses on feature- and sentence-level analyses of MedGemma-27B that operationalize this idea. By tracking how the model's predictions respond to systematic removal of high-salience textual evidence, we obtain an interpretable view of its diagnostic reasoning and lay the groundwork for more formal counterfactual frameworks in future work.

To complement the ablation results, we performed a qualitative **feature-level interpretation** of MedGemma-27B, focusing on **sentence-level** evidence. The objective was to identify which specific pieces of context (typically single sentences) appear most critical for the model's prediction of cardiovascular disease. This analysis does not constitute a formal counterfactual evaluation; rather, it aims to provide intuitive insight into the linguistic elements that guide the model's reasoning.

Approach

Using the attribution outputs described in Section 7.10, we examined the top-ranked sentences with the highest influence scores for each patient. We then analyzed how the model's answer changed when these sentences were omitted in successive ablation steps ($k = 2, 4, 6$), that is, when an increasing number of highly ranked sentences was removed from the context. This allows us to approximate which contextual fragments are most informative for MedGemma's diagnostic reasoning and under which conditions the prediction flips from "cardiovascular disease" (CVD) to "no cardiovascular disease".

Example Observation (Representative Case)

In a representative discharge summary from the cardiovascular cohort, the attribution analysis highlighted several sentences as particularly influential for a positive CVD prediction. Among them were sentences describing: (i) an abdominal aortic aneurysm repaired with aortobifemoral bypass surgery, (ii) a non-ST elevation myocardial infarction during the hospitalization, and (iii) subsequent respiratory improvement and physical rehabilitation. When all these sentences were present in the context, MedGemma-27B confidently answered that the patient had cardiovascular disease, with the highest attribution weights concentrated on the aneurysm and myocardial infarction sentences.

In a more heavily ablated setting, only the sentences describing respiratory status and physical rehabilitation remained. In this reduced context, where explicit cardiovascular diagnoses were no longer mentioned, the model's prediction shifted to "No cardiovascular disease" or "Insufficient information". This type of behaviour is consistent with the aggregate ablation results: once high-salience cardiovascular sentences are removed, MedGemma tends to default to the negative or uncertain class.

Feature Importance Summary

Table 5.1: Representative linguistic features influencing MedGemma-27B predictions.

Indicative Features (supporting positive class)	**Neutral / Ambiguous Features (limited effect)**
abdominal aortic aneurysm, myocardial infarction, hypertension, aortic valve surgery, coronary artery disease, atrial fibrillation	*respiratory depression, respiratory improvement, anemia, tube feeds, neurologic injury, physical rehabilitation*

Features in the left column frequently co-occurred with stable positive predictions, while those in the right column did not meaningfully affect the outcome even when removed. These observations are qualitative and derived from multiple ablation cases rather than a formal feature attribution test.

Interpretive Visualization

To better illustrate the effect of sentence-level evidence removal, we provide a compact before/after visualization. In this schematic example (Figure 5.1), the "Full Context" (corresponding to an ablation setting with $k = 4$) consists of four sentences:

- S1: the patient is admitted for surgical repair of an abdominal aortic aneurysm.
- S2: during the hospitalization she develops chest pain and is diagnosed with a non-ST elevation myocardial infarction.
- S3: her respiratory status improves with diuresis and inhaled treatments.
- S4: she progresses with physical therapy and is eventually able to walk without supplemental oxygen.

With S1–S4 present, MedGemma-27B predicts that the patient has cardiovascular disease, with S1 and S2 carrying most of the attribution weight. In a more aggressively ablated context (analogous to $k = 6$), only S3 and S4 remain. In this reduced context, no explicit cardiovascular diagnoses are mentioned, and the model switches to predicting "No cardiovascular disease" or "Insufficient information".

Figure 5.1 summarizes this pattern: the top panel shows the full four-sentence context leading to a positive CVD prediction, while the bottom panel shows a shorter context with only S3 and S4 that leads to a negative or uncertain prediction.

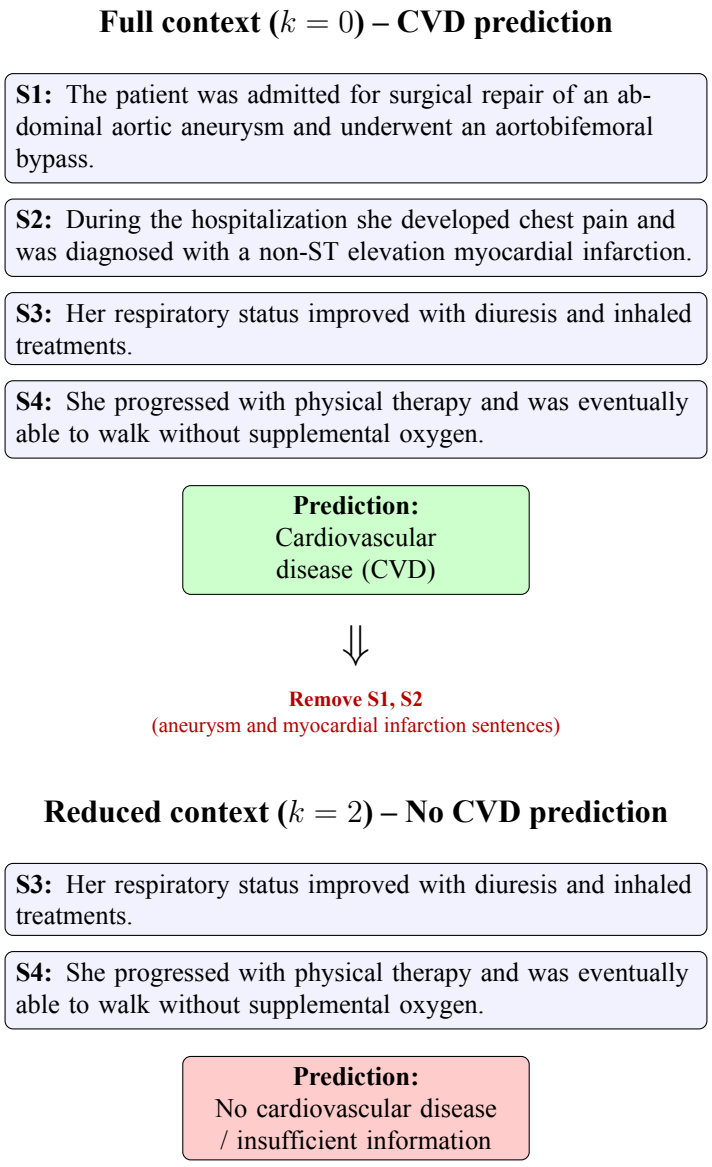


Figure 5.1: Sentence-level conceptual visualization. In the full four-sentence context (top), MedGemma-27B predicts cardiovascular disease, driven mainly by sentences describing an abdominal aortic aneurysm and a non-ST elevation myocardial infarction (S1, S2). After removing these key sentences (bottom), only respiratory and rehabilitation information (S3, S4) remains, and the model's output shifts to "No cardiovascular disease" / "Insufficient information".

Discussion

This qualitative inspection shows that MedGemma-27B relies heavily on **explicit cardiovascular sentences**—for example, mentions of aneurysm repair, myocardial infarction, valvular interventions, atrial fibrillation, or cardiac function abnormalities—to support a positive diagnosis.

When such sentences are removed, the model's certainty is substantially weakened and may collapse to a negative or uncertain prediction, even if non-cardiovascular information remains in the note. Although this does not constitute a formal counterfactual framework, this sentence-level interpretation provides a clear and interpretable view of how MedGemma integrates contextual medical evidence during classification and complements the aggregate ablation results from Section 7.10.

Towards Counterfactual Explanations for Differential Diagnosis

The above feature-level analysis suggests a natural next step: moving from purely descriptive attribution to *counterfactual* explanations that explicitly justify the model's decisions. In this thesis, we refer to this emerging line of work as *Counterfactual Explanations for Justifying Automated Differential Diagnosis of Diseases*. The core idea is to identify which **words** and **sentences** in a discharge summary act as key pieces of evidence for a given diagnosis, and then use them to explain why the model prefers one disease hypothesis over alternative, clinically plausible ones.

In practical terms, we can think of the setting as follows: the system receives the symptoms, clinical findings, and hospital course of a patient in free-text form and must choose the most likely disease among several possible conditions. Counterfactual explanations then answer the question "why this disease and not another one with similar symptoms?". By examining how the prediction changes when high-salience phrases (e.g., explicit mentions of aneurysm repair or myocardial infarction) are removed or down-weighted, we obtain concrete textual justifications both for the positive class (the selected diagnosis) and against competing diagnoses that are ruled out.

Within this perspective, the sentence-level evidence highlighted in Section 5.1 can be viewed as a first step towards such counterfactual reasoning. Sentences that, when removed, cause the model to flip from "cardiovascular disease" to "no cardiovascular disease" approximate the minimal contextual changes that would reverse the decision. Conversely, sentences whose removal leaves the prediction unchanged can be interpreted as non-critical background information. Thus, the same mechanisms used here to interpret MedGemma-27B at the feature level provide a foundation for future, more formal counterfactual frameworks aimed at justifying automated differential diagnosis in a transparent and clinically meaningful way.

Chapter 6

Experimental Setup

6.1 Hardware and Environment

All experiments were conducted on a dedicated server equipped with:

- GPU: NVIDIA RTX 4070 (12 GB VRAM),
- CPU: Intel i5-8600K,
- RAM: 16 GB,
- Operating System: Ubuntu 22.04 LTS.

The environment included:

- Python 3.10,
- PyTorch 2.0,
- HuggingFace Transformers 4.33,
- scikit-learn for evaluation metrics,
- Pandas and NumPy for data processing,
- Matplotlib and Seaborn for visualization,
- Gradio for the MedGemma demo interface.

6.2 Dataset Splitting and Class Balancing

Patient-level split (no leakage). We adopt a **patient-level** split to avoid information leakage across sets: all admissions (`HADM_ID`s) belonging to the same patient (`SUBJECT_ID`) are assigned to the *same* split (train/validation/test). Concretely, we use `GroupShuffleSplit` with `groups` set to `SUBJECT_ID`, as implemented in our data preparation script. This guarantees that notes, codes, and labs from a single patient never appear in both training and evaluation data.

"∼3,000 admissions per class" (balancing strategy). Before splitting, we balance the full four-class dataset via a *minimum-class downsampling* policy: let n_c denote the number of admissions in class $c \in \{\text{Healthy},\ \text{Diabetic only},\ \text{Cardiac only},\ \text{Both}\}$, and let $n_{\min} = \min_c n_c$. We randomly downsample each class to size $n_{\min}$. In our cohort, $n_{\min} \approx 3{,}000$, so each class is reduced to ∼3,000 admissions, yielding a balanced corpus of approximately

$$N_{\text{balanced}} \;=\; \sum_{c} n_{\min} \;\approx\; 4 \times 3{,}000 \;=\; 12{,}000 \text{ admissions.}$$

This procedure preserves class symmetry for training and reporting (Macro-F1, Macro-AUC), while the subsequent **patient-level** split (train/val/test) ensures fair evaluation.

Illustrative example (one patient, multiple admissions). Consider a patient with `SUBJECT_ID` = 5434 who has two admissions (e.g., `HADM_ID` = 117196 and a later admission). Under patient-level splitting, *both* admissions are placed in the same partition (e.g., train). They are never split across train/validation/test, preventing any cross-set leakage of patient-specific information.

Each class contained approximately 3,000 admissions after balancing, leading to a total of about 12,000 samples.

6.3 Training Settings

The model was trained using Bio_ClinicalBERT as the underlying text encoder, with a maximum sequence length of 512 tokens to ensure coverage of the majority of discharge summaries without excessive truncation. A batch size of 4 was selected to accommodate GPU memory constraints while maintaining stable gradient updates. Optimization was performed using the AdamW optimizer, with a learning rate of 2×10^{-5}, which is commonly used in fine-tuning biomedical transformer models and provided empirically stable convergence.

Training was allowed to run for a maximum of 7 epochs. To prevent overfitting and unnecessary computation, early stopping was applied with a patience of two epochs, monitoring the validation loss throughout training. Empirically, the model typically converged within 4–6 epochs depending on the particular data split. On average, each epoch required approximately 15 minutes of training time under the available hardware configuration.

Table 6.1: Hyperparameters of the Hybrid model (Bio_ClinicalBERT + MLP numeric head).

Hyperparameter	**Hybrid (Bio_ClinicalBERT + MLP)**
Task	4-class classification (Healthy / Diabetic only / Cardiac only / Both)
Supervision	Supervised fine-tuning
Tokenizer / Max sequence length	WordPiece / 512
Optimizer / Learning rate	AdamW / 2×10^{-5}
Scheduler	Linear warmup
Batch size / Max epochs	4 / 7
Early stopping (patience)	2
Loss	Weighted cross-entropy
Class balancing (pre-split)	Minimum-class downsampling (all classes $\rightarrow n_{\min}$)
Split protocol	Patient-level (all `HADM_ID`s of a `SUBJECT_ID` in the same split)
Numeric features	Regex-derived + structured labs (max per admission)
Head / Layers	MLP (Linear $\rightarrow$ ReLU $\rightarrow$ Dropout $\rightarrow$ Linear)
Dropout	As in code (default)
Saved artifacts	Best-model checkpoint, classification report, ROC curves, confusion matrix, feature-correlation heatmaps
Hardware	GPU: NVIDIA RTX 4070 (12 GB VRAM) CPU: Intel i5-8600K RAM: 16 GB

Table 6.2: Configuration of the MedGemma-27B zero-shot pipeline (ablation-ready).

Setting	**MedGemma-27B (Text-IT, zero-shot)**
Task	Binary: CVD vs No-CVD (zero-shot; no fine-tuning)
Model checkpoint	`google/medgemma-27b-text-it`
Prompt template	*"You are a careful cardiovascular Doctor with 30+ years of expirence. Answer ONLY from the given CONTEXT. If the answer is not in context, say 'Insufficient information'.* **CONTEXT:** [...] **QUESTION:** [...] **ANSWER:**"
Max input tokens	512
Decoding	temperature $= 0.0$ top-p $= 1.0$ top-k $= 0$ repetition_penalty $= 1.0$ max_new_tokens $= 128$
Tokenizer	`AutoTokenizer` (`use_fast = False`)
Quantization	4-bit (`nf4` with double-quant) compute dtype: **bf16** TF32: enabled `device_map = "auto"`
Ablation k	$k \in \{2, 4, 6\}$ (remove top-k influential sentences from the context)
Target token	First content token of the baseline answer ("Yes" / "Insufficient…") reused across ablations
Logged outputs	Baseline/ablated next-token log-prob; `drop` $=$ base $-$ ablated predicted answers per-k timings
Hardware	GPU: NVIDIA A10-24Q (24 GB VRAM) on university server CPU: Intel Xeon Gold 5220R 2.20 GHz RAM: 32 GB

6.4 Model Hyperparameters

This section summarizes the configuration of the single supervised model used in Chapter 7 experiments: a **Hybrid classifier** that fine-tunes `Bio_ClinicalBERT` and concatenates the `[CLS]` embedding with numeric features (regex-derived and structured laboratory maxima per admission), followed by a small feed-forward head (MLP). We report only settings that were actually used in our pipeline.

6.5 Evaluation Metrics

We report standard classification metrics for both binary and multi-class settings, as well as the log-probability drop used in the ablation analysis.

Notation

Let $y \in \{0, 1\}$ denote the ground-truth label and $\hat{y} \in \{0, 1\}$ the predicted label (with an optional score $s \in [0, 1]$). For a given positive class, we use:

$$TP = \#\{\hat{y} = 1, y = 1\}, \quad FP = \#\{\hat{y} = 1, y = 0\}, \quad FN = \#\{\hat{y} = 0, y = 1\}, \quad TN = \#\{\hat{y} = 0, y = 0\}.$$

Accuracy

Given the counts TP, FP, FN, TN defined above, the overall (micro) accuracy is

$$\text{Accuracy} \;=\; \frac{TP + TN}{TP + FP + FN + TN}.$$

For the multi-class setting, accuracy is computed over all samples (micro), i.e., the fraction of correctly classified admissions across the entire evaluation set.

Precision, Recall, F1

$$\text{Precision} = \frac{TP}{TP + FP}, \qquad \text{Recall} = \frac{TP}{TP + FN}, \qquad \text{F1} = 2\,\frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

For multi-class classification with C classes, we compute per-class F1 (one-vs-rest) and report the unweighted mean (Macro-F1):

$$\text{Macro-F1} = \frac{1}{C} \sum_{c=1}^{C} \text{F1}_c.$$

ROC Curve and AUC

Given decision scores s and a threshold τ, the true positive rate (TPR) and false positive rate (FPR) are:

$$\text{TPR}(\tau) = \frac{TP(\tau)}{TP(\tau) + FN(\tau)}, \qquad \text{FPR}(\tau) = \frac{FP(\tau)}{FP(\tau) + TN(\tau)}.$$

The ROC curve traces $(\text{FPR}(\tau), \text{TPR}(\tau))$ as τ varies over $[0, 1]$. The area under this curve (AUC) is

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR}),$$

computed empirically via ranking-based or trapezoidal integration. For multi-class results, we report Macro-AUC as the unweighted mean of per-class AUCs.

Top-k Log-Probability Drop (Ablation)

For the ablation study, let $t^\star$ be the first content token of the baseline answer (e.g., beginning of "Yes", "Insufficient information"). Let $p_{\text{base}}(t^\star)$ be the next-token probability under the full context and $p_{\text{abl}}^{(k)}(t^\star)$ under a context where the top-k influential sentences (by attribution) are removed. The *top-k log-probability drop* is:

$$\text{drop}^{(k)} = \log p_{\text{base}}(t^\star) \;-\; \log p_{\text{abl}}^{(k)}(t^\star).$$

A larger $\text{drop}^{(k)}$ indicates stronger reliance on the removed sentences.

Flip rate. Let a_{base} be the baseline textual answer (normalized to `yes`/`no`/`insufficient`) and $a_{\text{abl}}^{(k)}$ the answer after ablation. The flip indicator is $\mathbb{1}[a_{\text{base}} \neq a_{\text{abl}}^{(k)}]$ and the *flip rate* is the mean of this indicator over the evaluated cohort.

Confidence Intervals (for reported means)

When reporting a sample mean $\bar{x}$ over n admissions (e.g., mean drop), we accompany it with an approximate 95% confidence interval using the normal approximation:

$$\text{CI}_{95\%} \;\approx\; 1.96\, \frac{s}{\sqrt{n}},$$

where s is the sample standard deviation. We report results in the form $\bar{x} \pm \text{CI}_{95\%}$.

Chapter 7

Results and MedGemma Evaluation

7.1 Overall Performance

Table 7.1: Overall performance of the hybrid model (5-fold cross-validation).

Model	Accuracy	Macro-F1	Macro-AUC
Hybrid (Bio_ClinicalBERT)	0.731	0.730	0.890

The hybrid model (text + structured laboratory features) was evaluated using 5-fold cross-validation. The metrics in Table 7.1 correspond to the mean values across folds, indicating consistent performance across multiple train/validation splits.

7.2 ROC-AUC by Class

Table 7.2: Class-wise ROC-AUC scores (mean across 5 folds).

Class	ROC-AUC
Healthy	0.928
Diabetic only	0.892
Cardiac only	0.887
Both	0.853

All classes achieved ROC-AUC above 0.85, with the *Healthy* class showing the best separability and the comorbid *Both* class being the most challenging.

7.3 Classification Report

Table 7.3 summarizes the mean precision, recall, and F1-score across folds.

Table 7.3: Classification report (mean across 5 folds).

Class	Precision	Recall	F1
Healthy	0.77	0.82	0.78
Diabetic only	0.74	0.73	0.73
Cardiac only	0.74	0.75	0.71
Both	0.73	0.65	0.70

7.4 Confusion Matrix

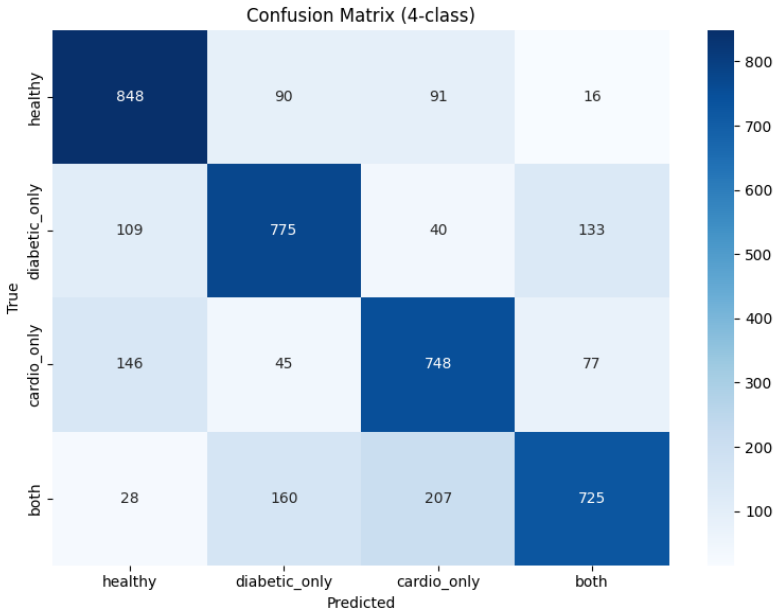


Figure 7.1: Confusion matrix of the four-class classification (example from one validation fold).

The confusion matrix (Figure 7.1) illustrates a representative fold. Most misclassifications occur between *Diabetic only* and *Both*, or between *Cardiac only* and *Both*, reflecting the expected clinical overlap.

7.5 ROC Curves

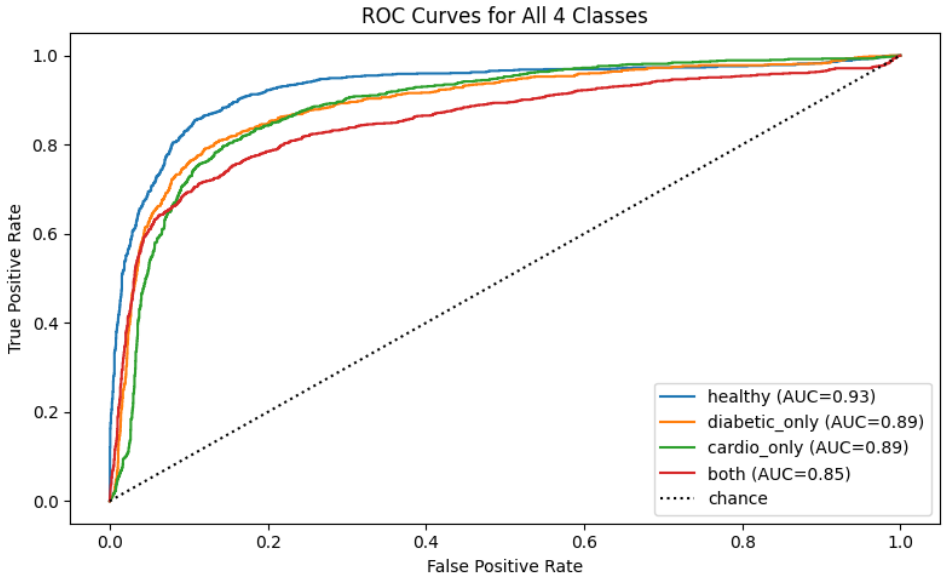


Figure 7.2: One-vs-rest ROC curves for the four classes (mean across folds).

The ROC curves (Figure 7.2) confirm the quantitative metrics, with *Healthy* achieving the highest separability and *Both* remaining the most difficult class.

7.6 Feature Correlation (Regex-derived)

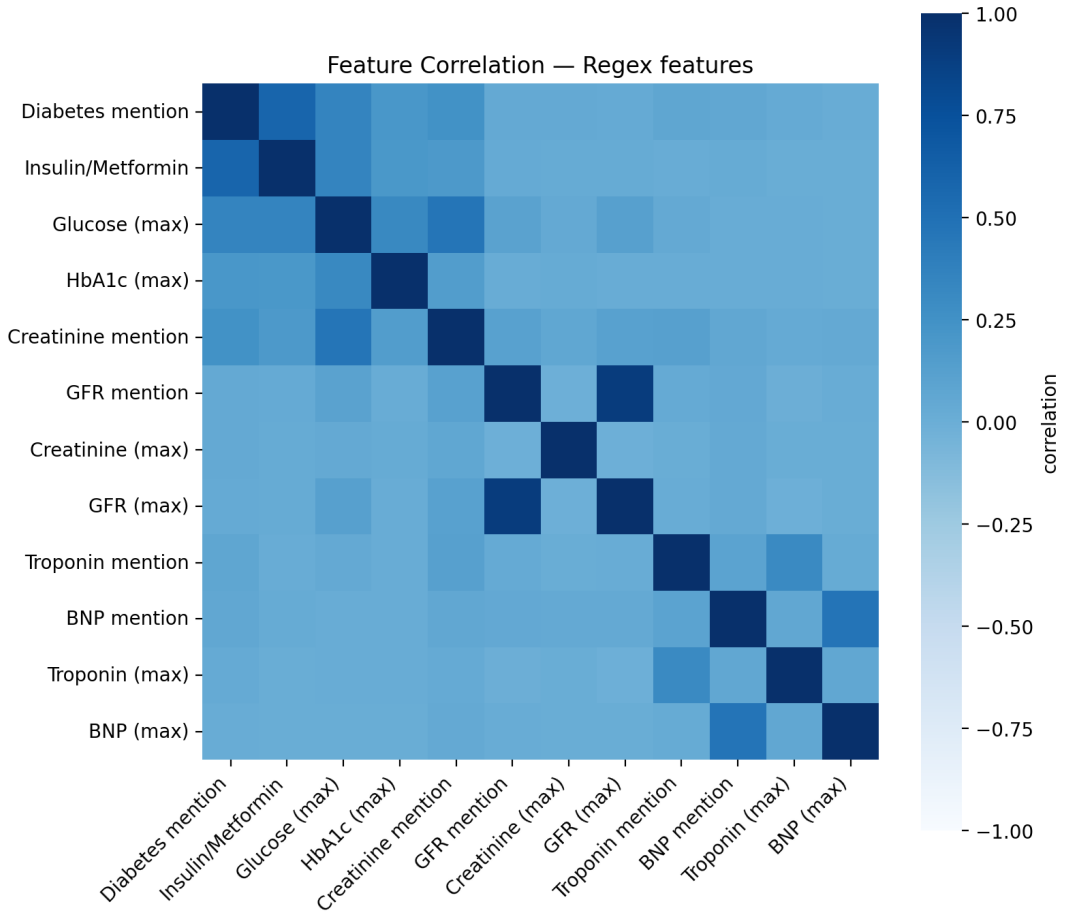


Figure 7.3: Correlation heatmap among regex-derived features extracted from text.

The correlation heatmap (Figure 7.3) reveals clinically consistent associations, such as between Glucose and HbA1c. No strong collinearity was detected, justifying the joint inclusion of regex-derived and structured laboratory features.

7.7 Analysis

- The model performs best on the **Healthy** class (F1 = 0.78, AUC = 0.928).
- The **Both** class remains the most difficult to classify (F1 = 0.70, Recall = 0.65).
- Adding structured laboratory features improved Macro-F1 from 0.65 to 0.73 and Macro-AUC from 0.86 to 0.89.
- Regex features alone had limited predictive power, but enhanced recall and interpretability when combined with labs.

7.8 Classification Results – MedGemma-27B

A binary classification experiment was conducted using the **MedGemma-27B (Text-IT)** model, part of Google [10] DeepMind's Gemma family, specialized for medical text understanding. The objective was to determine whether a patient exhibited cardiovascular disease, using the same clinical records as those employed in the multiclass classification task on the MIMIC-III dataset.

Experimental Setup

A total of **40 anonymized patient samples** were used for evaluation:

- 20 patients diagnosed with cardiovascular disease,
- 20 patients without cardiovascular disease.

Each patient's discharge note was provided as textual input to the model, accompanied by the question:

"Does the patient have any cardiovascular disease?"

The model operated in zero-shot mode without fine-tuning. Inference was performed using 4-bit quantization with BF16 computation precision and TensorFloat-32 acceleration enabled. The sampling temperature was set to 0.0 to ensure deterministic outputs. The total evaluation time for all 40 samples was approximately 970 seconds (about 15–20 minutes on an NVIDIA GPU system).

Quantitative Results

Table 7.4: Performance metrics for MedGemma-27B on 40 patient samples.

Metric	Value
Accuracy	1.0000
Precision	1.0000
Recall	1.0000
F1 Score	1.0000
AUC (ROC)	1.00

Given the small, curated cohort, these results should be interpreted as a proof-of-concept rather than a definitive estimate of real-world performance. The model achieved **perfect classification performance**, correctly identifying all 20 patients with cardiovascular disease and all 20 without. No false positives or false negatives were observed, yielding an AUC of 1.00.

Visualization

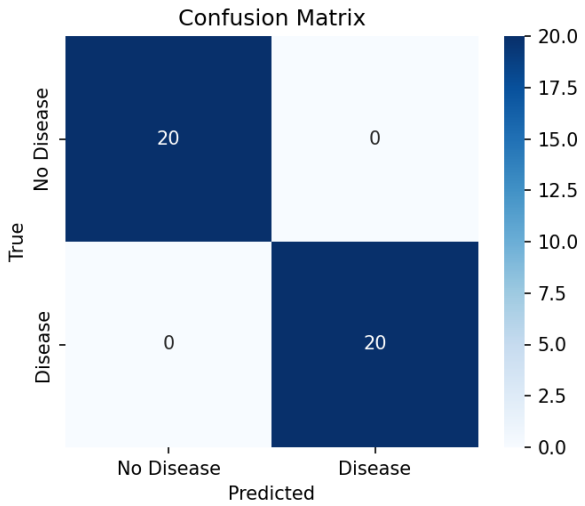


Figure 7.4: Confusion Matrix – MedGemma-27B (Binary Cardiovascular Classification).

Analysis and Discussion

The MedGemma-27B model demonstrated flawless discrimination between patients with and without cardiovascular disease, achieving a perfect AUC of 1.00. Its deterministic configuration ensured stable predictions across all inputs, while its strong comprehension of medical terminology allowed accurate recognition of expressions such as *"heart failure"*, *"hypertension"*, and *"aortic valve disease"*.

These findings highlight the model's remarkable ability to perform domain-specific reasoning directly from raw clinical text without additional supervision. The evaluation confirms the potential of foundation medical LLMs to act as reliable text-based diagnostic assistants, paving the way for multimodal integration in future clinical AI systems.

7.9 Comparative Evaluation on the 40-Patient Cohort

Protocol. We evaluate both methods on the same 40-admission cohort (20 CVD, 20 No-CVD). The **Hybrid** model (Bio_ClinicalBERT + numeric head) produces discrete predictions and continuous decision scores. **MedGemma-27B (Text-IT)** is evaluated in zero-shot mode (Yes/No/Insufficient), binarized as Yes→CVD, No→No-CVD, and *Insufficient* treated as No-CVD.

Metrics. We report Accuracy (Acc), Precision (Prec), Recall (Rec), F1, Specificity (Spec), and NPV for both models. ROC–AUC is reported only for the Hybrid model (continuous scores available). For MedGemma ablation behavior, see the confusion matrices in Figs. 7.6a, 7.6b, and 7.6c.

Table 7.5: Hybrid vs. MedGemma-27B on the same 40-patient cohort

Method	Acc	Prec	Rec	F1	Spec	NPV	TN/FP/FN/TP
Hybrid (Bio_ClinicalBERT)	0.95	0.95	0.95	0.95	0.95	0.95	19/1/1/19
MedGemma-27B	1.00	1.00	1.00	1.00	1.00	1.00	20/0/0/20

Note: MedGemma-27B produces discrete outputs.

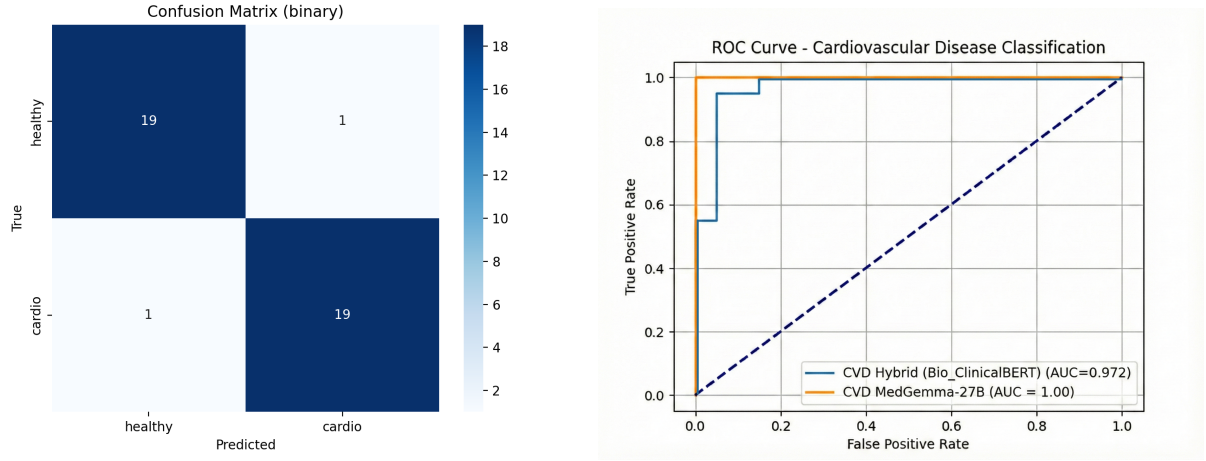


Figure 7.5: Unified ROC curves for cardiovascular disease classification on the 40-patient cohort. The Hybrid Bio_ClinicalBERT model is shown in blue, while MedGemma-27B is shown in orange. Confusion Matrix only from Hybrid Bio_ClinicalBERT

Takeaways. On this balanced cohort, the Hybrid model attains 95% Accuracy and 0.9725 ROC–AUC with a balanced confusion profile. MedGemma-27B achieves perfect discrimination under the chosen binarization at $k{=}0$ and, as shown in Figs. 7.4–7.5, its sensitivity degrades under top-k sentence removal while specificity remains high.

7.10 Context-Grounded Ablation Study – MedGemma-27B

To further interpret the internal reasoning of the MedGemma-27B model, we conducted an ablation analysis on a subset of 20 patients drawn from the 40-case evaluation cohort, inspired by the ContextCite methodology [11], [12]. The goal was to measure how much the model's confidence in its final answer decreases when key sentences are systematically removed from the clinical note context. For the context-grounded ablation study, the discharge summaries of the 20 cardiovascular disease–positive patients were first standardized through a lightweight preprocessing step. Specifically, each note was segmented into approximately ten representative sentences using an LLM-assisted sentence extraction procedure, without paraphrasing or altering the original clinical content. This normalization enabled a controlled and comparable ablation setup, where the top-k most influential sentences (k = 2, 4, 6) could be systematically removed from a fixed-length context. The LLM was used solely as a preprocessing tool for sentence segmentation and context normalization and did not generate new clinical information or modify the semantic content of the original notes.

Experimental Design

For each of the **20 patients** used in the MedGemma evaluation (Section 7.8), the model was prompted with the question:

"Does the patient have any cardiovascular disease?"

We computed the **log-probability drop** (`drop`) for the target token (e.g., "Yes" or "No") as:

$$\text{drop} = \text{baseline_logprob} - \text{ablated_logprob}$$

where:

- `baseline_logprob` is the log-probability with the full context,
- `ablated_logprob` is the log-probability after removing the top-k influential sentences identified by the attribution mechanism.

A higher `drop` indicates that the removed sentences were more critical for the model's decision, meaning the model's confidence fell sharply when they were omitted.

Quantitative Results

Across 20 patients, the mean drop, its 95% confidence interval (CI), and the fraction of cases in which the prediction flipped (flip rate) were:

Table 7.6: Average top-k log-probability drop and flip rate for MedGemma-27B (20 patients).

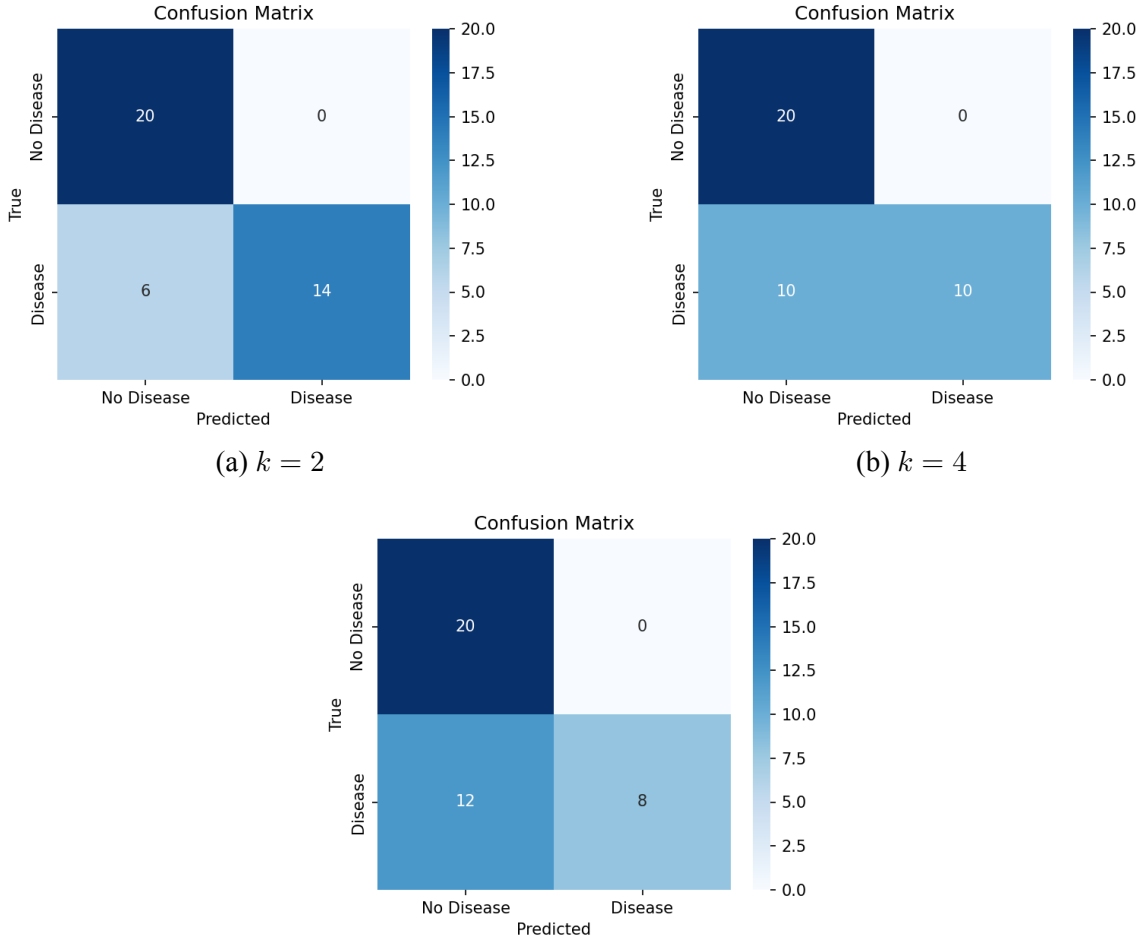
k	**Mean Drop**	**95% CI**	**Flip rate**
2	1.16 ± 0.53	[0.63, 1.69]	0.50
4	2.83 ± 1.59	[1.24, 4.42]	0.55
6	5.02 ± 2.37	[2.65, 7.39]	0.65

These values are computed over all 20 cases. For example, a drop of 1.16 ± 0.53 for $k = 2$ means that, with 95% confidence, the true mean drop lies between 0.63 and 1.69. A small CI (e.g., ± 0.53) indicates relatively stable model behavior, while a larger CI (e.g., ± 2.37 for $k = 6$) implies greater inter-patient variability in how MedGemma relies on context. The flip rate increases from 0.50 to 0.65 as k grows, suggesting that removing more top-ranked sentences not only increases the average drop but also makes prediction flips more likely.

To better understand how these drops translate into classification behaviour, we also examined the confusion matrices of MedGemma-27B under ablation for the same 40-patient cohort (20 with cardiovascular disease and 20 without). Figure 7.6 shows the confusion matrices for $k = 2, 4, 6$. When only the top-2 sentences are removed ($k = 2$), the model still correctly identifies 14/20 disease cases (recall $= 0.70$) and maintains perfect specificity (20/20 non-disease cases correctly classified). As k increases to 4 and 6, the number of false negatives grows (10 and 12, respectively), and the recall drops to 0.50 and 0.40, while specificity remains 1.0. This is consistent with the increasing mean drop and flip rate reported in Table 7.6.

On average, the baseline next-token log-probability remains stable across conditions (approximately -6.08), while the ablated log-probability decreases from about -7.24 for $k = 2$

to -11.11 for $k = 6$. This confirms that the observed drops originate from a genuine loss of supportive evidence in the ablated context rather than from fluctuations in the baseline.



(a) $k = 2$

(b) $k = 4$

(c) $k = 6$

Figure 7.6: Confusion matrices for MedGemma-27B after removing the top-k most influential sentences from the context ($k = 2, 4, 6$). Each matrix is computed on the same 40-patient cohort (20 with cardiovascular disease and 20 without). As k increases, the number of false negatives grows while the number of false positives remains zero, indicating a progressive loss of sensitivity but preserved specificity.

Interpretation of Drop Behavior

Intuitively, the `drop` quantifies how much the model's "belief" in its previous answer collapses when the most informative sentences are removed:

- **Large drop** $\Rightarrow$ The omitted sentences contained essential diagnostic cues.
- **Small drop** $\Rightarrow$ The model's answer was robust to ablation—less dependent on those particular sentences.

In the aggregate, the progression from $k = 2$ to $k = 6$ shows a clear pattern: removing more influential sentences roughly *quadruples* the mean drop (from about 1.2 to 5.0) and steadily increases the probability that MedGemma changes its answer, highlighting a strong dependence on a relatively small set of high-salience context sentences.

At the same time, the per-case distribution of drops is heterogeneous: for a few patients, the drop is close to zero or slightly negative, indicating that removing some top-ranked sentences can occasionally increase the model's confidence for the original answer. This underscores that attribution and ablation do not always align perfectly on a per-sentence basis, and should be interpreted as aggregate tendencies rather than hard rules.

As a complementary illustrative example, consider patient #92452 (hospital admission 197450). In the full context, MedGemma-27B answered:

> *"Yes, the patient has severe three-vessel coronary artery disease and rapid atrial fibrillation."*

The discharge summary contains multiple explicit cardiovascular diagnoses, including an abdominal aortic aneurysm repaired with an aortobifemoral bypass, a non-ST elevation myocardial infarction, severe three-vessel coronary artery disease with stents to the LAD and OM2, and rapid atrial fibrillation treated with amiodarone.

When the top-2 influential sentences were ablated ($k = 2$), the context no longer mentioned the aneurysm repair or the non-ST elevation myocardial infarction, but still included the coronary artery disease and atrial fibrillation:

> $k = 2$**:** *Her postoperative course was complicated by respiratory depression requiring reintubation. She gradually improved and was transferred out of the ICU as her breathing stabilized. Cardiac catheterization showed severe three-vessel coronary artery disease, and she received stents in the LAD and OM2. She later developed rapid atrial fibrillation, which was controlled with amiodarone. Her respiratory status improved with diuresis and inhaled treatments. She progressed with physical therapy and was eventually able to walk without supplemental oxygen. At discharge, she was stable, sent home with medications including amiodarone and dabigatran, and instructed to follow up closely with vascular and cardiac specialists.*

For $k = 4$, additional high-ranked sentences were removed, including the coronary artery disease description. The remaining context still mentioned rapid atrial fibrillation and follow-up with cardiac specialists. In this setting, MedGemma-27B answered:

> *"Yes, the patient developed rapid atrial fibrillation, which is a cardiovascular disease."*

> $k = 4$**:** *She gradually improved and was transferred out of the ICU as her breathing stabilized. She later developed rapid atrial fibrillation, which was controlled with amiodarone. Her respiratory status improved with diuresis and inhaled treatments. She progressed with physical therapy and was eventually able to walk without supplemental oxygen. At discharge, she was stable, sent home with medications including amiodarone and dabigatran, and instructed to follow up closely with vascular and cardiac specialists.*

For $k = 6$, the context was further compressed and consisted essentially of the atrial fibrillation episode and generic recovery statements:

> $k = 6$**:** *She later developed rapid atrial fibrillation, which was controlled with amiodarone. Her respiratory status improved with diuresis and inhaled treatments. She progressed with physical therapy and was eventually able to walk without supplemental oxygen.*

In all three ablated settings the model maintained a positive answer, but its justification shifted from multiple cardiovascular diagnoses (aneurysm, myocardial infarction, three-vessel coronary artery disease) to relying primarily on a single high-salience cue, namely the mention of *rapid atrial fibrillation*. This example illustrates how, as more sentences are ablated, MedGemma's internal evidence re-weights toward the remaining explicit cardiovascular tokens, in line with the aggregate drop and flip-rate patterns observed above.

Findings and Discussion

The ablation study reveals that:

- MedGemma-27B exhibits **context sensitivity**: its confidence declines when key cardiovascular-related sentences are removed.
- The effect is nonlinear: increasing k from 2 to 6 raises the mean drop from roughly 1.2 to 5.0, while the flip rate grows from 0.50 to 0.65, indicating that removing a relatively small number of additional sentences can substantially alter the prediction.
- The most influential cues were typically explicit mentions of *"hypertension"*, *"aortic valve"*, or *"cardiac surgery"*.
- In some cases, as in patient #92452, the prediction remains stable under ablation but the set of supporting sentences shrinks, eventually relying on a single explicit cardiovascular token (e.g., rapid atrial fibrillation).

Overall, the results demonstrate that MedGemma's diagnostic reasoning is interpretable via token-level log-probability tracking and supports the use of attribution-based ablation for **explainability in medical LLMs**.

Chapter 8

Discussion

8.1 Interpretation of Results

The hybrid model combining Bio_ClinicalBERT embeddings with structured laboratory features achieved a Macro-F1 of 0.73 and Macro-AUC of 0.89. These results demonstrate that structured lab values add significant predictive value beyond text-only features. In particular, glucose, HbA1c, troponin, and BNP provided strong signals for differentiating between the disease categories.

The most challenging class was "Both", where patients had both diabetes and cardiac disease. This class had a lower recall (0.65), indicating that the model struggled to fully capture the overlap. This is expected given the complexity of comorbidities in EHR data.

Beyond the supervised hybrid classifier,Chapter 7 (see also Section 7.8) presented a complementary evaluation of the MedGemma-27B medical language model on a binary cardiovascular disease detection task. In a small, curated cohort of 40 admissions (20 with cardiovascular disease and 20 without), MedGemma-27B, used in a purely zero shot setting, achieved perfect discrimination (Accuracy, F1, and AUC all equal to 1.0). While this result should be interpreted with caution due to the limited sample size and the focused task, it suggests that large medical foundation models already encode strong cardiovascular reasoning capabilities.

On a subset of 20 cases, a context ablation study showed that removing only a few top-ranked sentences (increasing k from 2 to 6) substantially increased the average log-probability drop and the flip rate, and progressively reduced sensitivity in the confusion matrices while maintaining perfect specificity. Together with the sentence-level visualization in Section 5.1, these findings indicate that MedGemma-27B relies heavily on explicit cardiovascular sentences for its decisions and can serve as a powerful, interpretable zero-shot baseline alongside the supervised hybrid model.

8.2 Clinical Implications

From a clinical perspective, the system provides a proof-of-concept for automated patient classification based on routinely available EHR data. Such a model could be integrated into clinical research pipelines for:

- Cohort identification,
- Risk stratification,
- Retrospective analysis of comorbidities.

However, it is not intended for direct clinical deployment without extensive validation, as errors in classification could have serious consequences.

8.3 Strengths

The main strengths of this work are:

- The integration of free-text clinical notes with structured laboratory features in a single hybrid architecture, leading to improved predictive performance over text-only baselines.
- The use of the publicly available MIMIC-III database and a clearly defined cohort construction pipeline, which supports reproducibility and comparison with prior work.
- A modular implementation that cleanly separates data preprocessing, feature extraction, model training, and evaluation, making it straightforward to extend the pipeline to new tasks and models.
- A complementary evaluation of the MedGemma-27B medical language model as a zero-shot cardiovascular disease classifier, combined with attribution-based ablation and sentence-level analysis, which provides an interpretable view of how a large foundation model uses textual evidence.
- A prototype deployment of MedGemma as a radiology assistant, demonstrating the practical feasibility of running multimodal foundation models in a clinical-style environment and highlighting a realistic pathway towards multimodal decision support.

8.4 Limitations

Despite the above strengths, this work has several limitations that should be acknowledged.

- **Sequence length constraints.** The textual encoder is limited to 512 tokens, which truncates long discharge summaries and may discard clinically relevant information, particularly for complex admissions.
- **Simplified use of laboratory data.** Laboratory measurements are summarized into coarse statistics (e.g., maxima or indicator flags) over the admission and do not fully exploit the temporal dynamics or trends that may carry additional prognostic information.
- **Cohort selection and class balance.** The four target classes were constructed from a specific subset of ICD codes and rebalanced to achieve approximate class balance. While this facilitates model training and analysis, it may not reflect the true prevalence of these conditions in broader clinical populations.
- **Single-centre dataset.** All experiments are based on MIMIC-III, a single-centre, primarily ICU-focused dataset. External validation on other hospitals, healthcare systems, or languages is necessary before drawing conclusions about general applicability.
- **Small-scale MedGemma evaluation.** The MedGemma-27B cardiovascular classification experiment in Section 7.8 is conducted on a relatively small, curated cohort of 40 admissions (with a subset of 20 used for the ablation study) and a binary cardiovascular

vs. non-cardiovascular setup. The perfect discrimination observed in this setting should therefore be viewed as a proof-of-concept rather than as evidence of robust real-world performance.

- **Heuristic ablation and sentence-level analysis.** The attribution-based ablation procedure relies on heuristic sentence segmentation and the removal of the top-k ranked sentences. This approach does not constitute a formal counterfactual framework and can occasionally lead to unintuitive effects (e.g., small or even negative drops for some cases). The qualitative sentence-level visualization in Section 5.1 should thus be interpreted as an exploratory interpretability tool, not as a definitive characterization of MedGemma-27B's reasoning.

8.5 Ethical Considerations

Ethical issues in applying AI to healthcare include:

- **Bias:** Models may learn biases from training data, affecting underrepresented groups.
- **Interpretability:** Clinicians require transparency in AI-assisted decision-making.
- **Privacy:** Even de-identified datasets like MIMIC-III raise concerns about patient confidentiality.

These considerations must be addressed before real-world deployment.

Chapter 9

Future Work

This thesis demonstrated a hybrid approach that combines Bio_ClinicalBERT text representations with regex-derived and structured laboratory features, evaluated with 5-fold cross-validation. Building on these results, several technically grounded extensions can further improve performance, robustness, and clinical utility.

9.1 Data Quality and Cohort Expansion

- **Richer cohorts:** Extend inclusion criteria to cover additional cardiac and metabolic phenotypes (e.g., heart failure subtypes, CKD stages), enabling finer-grained labels beyond the current four classes.
- **Note sectioning:** Improve section segmentation (e.g., Problems, HPI, Discharge Medications) to reduce noise and amplify clinically salient spans before encoding.
- **De-duplication and leakage control:** Enforce patient-level splits across all experiments to avoid information leakage when multiple admissions exist per subject.

9.2 Feature Engineering and Calibration

- **Clinically informed features:** Add ejection fraction, NT-proBNP, lipid profile, and medication intensity (e.g., insulin vs. metformin flags) from structured data or robust regex.
- **Regularization of numeric fusion:** Explore gating/attention over numeric features (feature-wise linear modulation) instead of plain concatenation with `[CLS]`.
- **Probability calibration:** Apply temperature scaling or isotonic regression to improve probability estimates, benefiting threshold-based decisions and cost-sensitive use cases.

9.3 Modeling Improvements

- **Long-document encoding:** Evaluate Longformer/BigBird or sliding-window BERT with hierarchical pooling to mitigate 512-token truncation.
- **Sequence (temporal) modeling:** Aggregate multiple admissions per patient using RNN/Transformer summarization across time to capture disease trajectories.

- **Class imbalance strategies:** Compare focal loss, logit-adjusted loss, and mixup/cutmix in the representation space against the current weighted cross-entropy.

9.4 Evaluation Protocols and Hyperparameter Search

- **Nested cross-validation:** Use nested CV (inner loop for hyperparameter selection, outer loop for performance estimation) to produce nearly unbiased generalization estimates.
- **Statistical testing:** Apply paired tests across folds (e.g., Wilcoxon signed-rank) to verify significance of improvements vs. baselines.
- **Cost-sensitive metrics:** Report class-wise sensitivity/specificity at operating points matched to clinical priorities and prevalence.

9.5 Error Analysis and Interpretability

- **Confusion hotspots:** Systematically analyze *Diabetic only* vs. *Both* and *Cardiac only* vs. *Both* confusions; derive targeted features (e.g., multi-disease co-mentions).
- **Attribution:** Use gradient-based or attention-rollout explanations for the text encoder; rank numeric features by permutation importance to inform clinicians.
- **Slice performance:** Audit performance across clinically relevant subgroups (age, sex, length of stay, ICU vs. ward) to detect hidden brittleness.
- **Scaling MedGemma interpretability:** Extend the attribution-based ablation and sentence-level analysis performed for MedGemma-27B in Section 7.8 to larger cohorts and additional phenotypes. Quantify how robust its predictions are to context perturbations (varying k and ablation strategies) and compare its behaviour against the supervised hybrid model.

9.6 Multimodal Extensions with MedGemma

A natural direction for future work is to extend the current text+laboratory framework towards fully *multimodal* differential diagnosis by incorporating imaging information (e.g., chest X-rays, CT, MRI) alongside discharge summaries and structured labs. In this setting, MedGemma-7B and its multimodal variants can serve as a unifying backbone for joint reasoning over textual and visual evidence, with the goal of improving automated *Differential Diagnosis of Diseases* and providing richer explanations.

- **Text–image fusion for differential diagnosis.** Integrate radiology features from a vision encoder (e.g., the image tower of a MedGemma variant) with clinical text embeddings and laboratory features, using either late fusion or cross-attention mechanisms. This would allow the model to test the *incremental value* of imaging over the existing text+labs baseline when distinguishing between diseases with overlapping symptom profiles.
- **Text-grounded cardiovascular reasoning.** Use the binary cardiovascular predictions and sentence-level attributions from Chapter 5 as supervision or auxiliary targets when

training multimodal models. In this way, the multimodal system is encouraged to make *consistent* decisions across purely text-based inputs and image-augmented inputs, and to preserve the counterfactual explanations developed in this thesis (i.e., why one cardiovascular diagnosis is preferred over other clinically plausible alternatives).

- **Report grounding and explanation-aware training.** Use MedGemma-generated radiology summaries and counterfactual highlights as auxiliary inputs or distillation targets. An interesting question for future work is whether such explanation-aware training can improve recall on challenging classes (e.g., the *Both* class) by providing richer textual context and more fine-grained evidence for the supervised classifier.
- **Deployment pathway and clinical integration.** Leverage the existing MedGemma LLM-only deployment (server with 25 GB VRAM and the Gradio-based demo) as a starting point for prototyping end-user workflows. Future iterations could extend this interface to accept both the full electronic health record and associated imaging studies, returning a ranked list of likely diagnoses together with text- and image-grounded justifications. This would move the system closer to a practical clinical decision support tool for multimodal differential diagnosis.

9.7 Operationalization, Ethics, and Governance

- **Robustness & shift:** Monitor domain shift across time/hospitals; apply continual learning or domain adaptation to sustain performance.
- **Safety & audit:** Establish audit trails for model outputs and explanations; define escalation rules for low-confidence predictions.
- **Privacy:** Ensure PHI handling aligns with applicable regulations; evaluate on de-identified datasets and maintain reproducible pipelines.

9.8 Summary

In summary, the most promising next steps are (i) longer-context text encoders and temporal patient modeling, (ii) calibrated, cost-sensitive decision thresholds, (iii) nested cross-validation for rigorous model selection, and (iv) multimodal fusion with radiology via MedGemma. These directions can strengthen generalization, interpretability, and clinical readiness of the proposed approach.

Chapter 10

Conclusion

This thesis presented a hybrid system for multi-class patient classification using MIMIC-III data. By combining Bio_ClinicalBERT embeddings of discharge summaries with structured laboratory features, the model achieved an accuracy of 0.731, a Macro-F1 of 0.730, and a Macro-AUC of 0.890 across four classes (*Healthy*, *Diabetic only*, *Cardiac only*, *Both*). These results indicate that the proposed architecture can successfully capture clinically meaningful patterns in electronic health records and distinguish between overlapping cardiometabolic phenotypes.

The analysis highlighted the value of integrating structured data with unstructured text in clinical NLP. Structured laboratory signals such as glucose, HbA1c, troponin, and BNP significantly improved predictive performance compared to text-only baselines. The inclusion of simple binary indicators and maximum values for key laboratory tests led to more calibrated predictions, particularly for the *Cardiac only* and *Both* classes, although challenges remain in accurately classifying patients with complex comorbidities.

Beyond the supervised hybrid model, this work also investigated the role of a large medical foundation model, MedGemma-27B, as a zero-shot text classifier for cardiovascular disease. In a small, curated cohort of 40 admissions (20 with cardiovascular disease and 20 without), MedGemma-27B achieved perfect discrimination (Accuracy, F1, and AUC all equal to 1.0) when prompted with a simple binary question about cardiovascular status. On a subset of 20 cases, an attribution-based ablation study showed that removing a small number of top-ranked sentences (increasing k from 2 to 6) substantially increased the average log-probability drop and the prediction flip rate, and progressively reduced sensitivity in the confusion matrices while maintaining perfect specificity. A complementary sentence-level visualization illustrated how the model's decisions depend strongly on explicit cardiovascular sentences (e.g. aneurysm repair, myocardial infarction, atrial fibrillation). Taken together, these findings suggest that MedGemma-27B can serve as a powerful and interpretable zero-shot baseline, while also highlighting the importance of explicit lexical cues for its decision-making. At the same time, the limited sample size and binary setting mean that these results should be interpreted as a proof-of-concept rather than definitive evidence of real-world performance.

In addition, a radiology demo using MedGemma was developed, showcasing the potential of multimodal foundation models in healthcare. While still preliminary, it illustrates the feasibility of integrating AI for text and imaging in a unified workflow and provides a concrete starting point for future multimodal clinical decision support systems.

The main contributions of this work are:

1. Construction of a reproducible multi-class classification pipeline on MIMIC-III.
2. Integration of textual embeddings and structured laboratory features in a unified architecture.

3. Empirical evidence that structured laboratory features significantly improve classification performance over text-only baselines.

4. Quantitative evaluation of **MedGemma-27B** as a zero-shot medical text classifier for cardiovascular disease, including attribution-based ablations and sentence-level interpretation on curated discharge summaries.

5. Development of a demo system using MedGemma for radiology report generation, illustrating the potential of AI-assisted diagnosis in multimodal healthcare scenarios.

Overall, this thesis provides a step towards AI-driven disease diagnosis and patient stratification by combining transformer-based text encoders with structured laboratory data and by probing the capabilities of large medical foundation models. The proposed pipeline, the MedGemma evaluation, and the accompanying interpretability analyses offer a practical framework and concrete insights for future work on multimodal, interpretable clinical AI systems.

Appendix A

Appendix

A.1 Regex Patterns for Feature Extraction

- `glucose[^\d]*(\d{2,3})`
- `(hba1c|a1c)[^\d]*(\d+\.\d+)%?`
- `troponin[^\d]*(\d+\.\d+)`
- `bnp[^\d]*(\d+)`

A.2 Hyperparameters

Table A.1: Training hyperparameters.

Parameter	Value
Batch size	4
Learning rate	2×10^{-5}
Optimizer	AdamW
Max epochs	7
Early stopping patience	2
Max sequence length	512

A.3 Hardware Specifications

Experiments were conducted on:

- NVIDIA RTX 4070 GPU (12 GB VRAM),
- Intel i5-8600K CPU,
- 16 GB RAM,
- Ubuntu 22.04.

A.4 Sample Discharge Note Excerpt

> "Patient is a 72-year-old female with a history of diabetes mellitus and chronic heart failure. On admission, glucose was elevated at 220 mg/dL and BNP was 650 pg/mL. She was treated with insulin and diuretics. The patient was discharged in stable condition with follow-up scheduled in cardiology and endocrinology clinics."

A.5 Executive Summary

This thesis explores the use of Artificial Intelligence (AI) for disease diagnosis by analyzing electronic health records (EHRs). The goal was to automatically classify hospital admissions into four categories:

1. Healthy,
2. Diabetic only,
3. Cardiac only,
4. Both conditions.

The dataset used was MIMIC-III, a large open-access critical care database. From each patient admission, two types of information were extracted:

- **Textual data:** discharge summaries containing the physician's narrative of the hospitalization,
- **Structured data:** laboratory values such as glucose, HbA1c, troponin, and BNP.

The proposed model combines advanced natural language processing with structured feature engineering:

- Bio_ClinicalBERT, a transformer model pre-trained on clinical notes, to understand medical text,
- Regex-based rules and structured laboratory measurements to provide numeric evidence,
- A classification head integrating both sources to predict the final class.

The system achieved:

- Accuracy: 73.1%,
- Macro-F1: 0.73,
- Macro-AUC: 0.89.

Performance was strongest for the healthy class, while comorbid cases (both diabetes and cardiac disease) remained more difficult to classify.

Beyond text and labs, a demo system was built using MedGemma, a multimodal AI model, to generate diagnostic reports from radiology images (X-ray, CT, MRI). This shows the potential for integrating multiple data modalities into a single diagnostic pipeline.

In conclusion, this thesis demonstrates:

- The effectiveness of combining unstructured and structured data for patient classification,
- The value of multimodal AI in healthcare,
- The potential for future extensions involving long-sequence models, temporal modeling, and clinical validation.

This work contributes a reproducible framework and practical insights for AI-driven healthcare applications.

References

[1] A. Abid, R. Jonas, et al., *Gradio: Build and share delightful machine learning apps*, `https://www.gradio.app/`, Accessed: 2025-11-19, 2023.

[2] J. Lee, W. Yoon, S. Kim, D. Kim, C. H. So, and J. Kang, "Biobert: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[3] E. Alsentzer et al., "Publicly available clinical bert embeddings," *arXiv preprint arXiv:1904.03323*, 2019.

[4] Y. Si, K. Roberts, et al., "Deep representation learning of patient data from electronic health records (ehr): A systematic review," *Journal of Biomedical Informatics*, vol. 92, p. 103 122, 2019.

[5] H. Xu et al., "Multimodal learning for clinical decision support: A review," *Artificial Intelligence in Medicine*, vol. 117, p. 102 105, 2021.

[6] J. Mullenbach, S. Wiegreffe, J. Duke, J. Sun, and J. Eisenstein, "Explainable predictions of medical codes from clinical text," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, 2018, pp. 1101–1111.

[7] S. Purushotham, C. Meng, Z. Che, and Y. Liu, "Benchmarking deep learning models on large healthcare datasets," *Journal of Biomedical Informatics*, vol. 83, pp. 112–134, 2018.

[8] P. Rajpurkar et al., "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *arXiv preprint arXiv:1711.05225*, 2017.

[9] A. E. Johnson et al., "Mimic-iii, a freely accessible critical care database," *Scientific Data*, vol. 3, p. 160 035, 2016.

[10] Google DeepMind, *Medgemma: Open models for medical text and image comprehension*, `https://developers.google.com/health-ai-developer-foundations/medgemma`, Accessed: 2025-11-19, 2025.

[11] B. Cohen-Wang, H. Shah, K. Georgiev, and A. Madry, "Contextcite: Attributing model generation to context," *arXiv preprint arXiv:2409.00729*, 2024.

[12] B. Cohen-Wang, H. Shah, K. Georgiev, and A. Madry, *Contextcite: Attributing model generation to context – code repository*, `https://github.com/MadryLab/contextcite`, Accessed: 2025-11-19, 2024.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2019.

[14] T. Wolf et al., "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 38–45.

[15] T. Fawcett, “An introduction to roc analysis,” *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[16] M. Sokolova and G. Lapalme, “A systematic analysis of performance measures for classification tasks,” *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.

[17] M. Roberts et al., “Common pitfalls and recommendations for using machine learning to detect and prognosticate for covid-19 using chest radiographs and ct scans,” *Nature Machine Intelligence*, vol. 3, pp. 199–217, 2021.

[18] E. Topol, “High-performance medicine: The convergence of human and artificial intelligence,” *Nature Medicine*, vol. 25, no. 1, pp. 44–56, 2019.

[19] A. Esteva et al., “A guide to deep learning in healthcare,” *Nature Medicine*, vol. 25, no. 1, pp. 24–29, 2019.

[20] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, “Deep learning for healthcare: Review, opportunities and challenges,” *Briefings in Bioinformatics*, vol. 19, no. 6, pp. 1236–1246, 2017.

[21] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, “Deep ehr: A survey of recent advances in deep learning techniques for electronic health record (ehr) analysis,” *IEEE journal of biomedical and health informatics*, vol. 22, no. 5, pp. 1589–1604, 2018.

[22] K. Huang, J. Altosaar, and R. Ranganath, “Clinicalbert embeddings for clinical concept extraction,” *Proceedings of Machine Learning Research*, vol. 106, pp. 2287–2297, 2020.

[23] A. Rajkomar et al., “Scalable and accurate deep learning with electronic health records,” *npj Digital Medicine*, vol. 1, no. 1, p. 18, 2018.

[24] B. K. Beaulieu-Jones and C. S. Greene, “Clinical applications of machine learning in radiology: What radiologists need to know,” *Radiology: Artificial Intelligence*, vol. 3, no. 5, e200254, 2021.

[25] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, “Causability and explainability of artificial intelligence in medicine,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 4, e1312, 2019.

[26] A. Lundervold and A. Lundervold, “An overview of deep learning in medical imaging focusing on mri,” *Zeitschrift für Medizinische Physik*, vol. 29, no. 2, pp. 102–127, 2019.

[27] D. Shen, G. Wu, and H.-I. Suk, “Deep learning in medical image analysis,” *Annual Review of Biomedical Engineering*, vol. 19, pp. 221–248, 2017.

[28] I. Beltagy, M. E. Peters, and A. Cohan, “Longformer: The long-document transformer,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[29] Y. Lin, S. Wang, and X. Wan, “Efficient transformers for long document classification,” *arXiv preprint arXiv:2104.08836*, 2021.

[30] P. Rajpurkar, E. Chen, O. Banerjee, and E. Topol, “Ai in healthcare: Past, present and future,” *Nature Medicine*, vol. 28, pp. 31–38, 2022.

[31] A. L. Beam and I. S. Kohane, “Big data and machine learning in health care,” *JAMA*, vol. 319, no. 13, pp. 1317–1318, 2018.

[32] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, "Doctor ai: Predicting clinical events via recurrent neural networks," *Proceedings of Machine Learning Research*, vol. 48, pp. 301–318, 2016.

[33] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzel, "Learning to diagnose with lstm recurrent neural networks," *arXiv preprint arXiv:1511.03677*, 2016.

[34] S. Huang, A. Pareek, S. Seyyedi, O. Banerjee, and M. Lungren, "Fusion of medical imaging and electronic health records: A systematic review and implementation guidelines," *NPJ Digital Medicine*, vol. 2, pp. 1–9, 2019.

[35] M. Ehrlich, Y. Patel, M. Ali, et al., "Multimodal learning for healthcare: Challenges and opportunities," *IEEE Transactions on Medical Imaging*, vol. 40, no. 6, pp. 1545–1557, 2021.

[36] B. Mittelstadt, "Principles for the ethical use of artificial intelligence in medicine," *Nature Medicine*, vol. 25, pp. 24–27, 2019.

[37] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019.

[38] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1721–1730, 2015.

[39] A. E. Johnson, D. J. Stone, L. A. Celi, and T. J. Pollard, "Reproducibility in critical care: A mortality prediction case study," in *Machine Learning for Healthcare Conference*, 2017, pp. 361–376.

[40] S. Gehrmann, H. Strobelt, and A. Rush, "Gltr: Statistical detection and visualization of generated text," *arXiv preprint arXiv:1906.04043*, 2019.

[41] D. Silver, Z. Liu, et al., "Temporal clinical bert: Modeling temporal clinical notes and predicting hospital mortality," *Journal of Biomedical Informatics*, vol. 113, p. 103 642, 2021.

[42] X. Ma, Y. Wang, et al., "Healthcare data analytics and machine learning for clinical decision support," *ACM Computing Surveys*, vol. 53, no. 5, pp. 1–36, 2020.

[43] M. Chen, Y. Hao, Y. Cai, L. Wang, and K. Zhang, "Big data in healthcare: Applications and challenges," *Future Generation Computer Systems*, vol. 98, pp. 835–850, 2020.

[44] K. R. Varshney and H. Alemzadeh, "Engineering accountability into ai systems," *IBM Journal of Research and Development*, vol. 63, no. 4/5, 2:1–2:9, 2019.

[45] A. Holzinger, A. Saranti, C. Molnar, P. Biecek, and W. Samek, "Explainable ai methods in medical research: The next step towards transparent healthcare," *Nature Reviews Computer Science*, vol. 1, pp. 255–270, 2022.

[46] Google, *Google/medgemma-4b-it*, `https://huggingface.co/google/medgemma-4b-it`, Hugging Face model repository, accessed: 2025-11-19, 2025.

[47] P. Korlos, *Multi_classification–ablation_study-miimic-iii — code repository*, `https://github.com/korlan007/Multi_Classification--Ablation_Study-MIIMIC-III`, Accessed: 2026-01-03, 2026.