School of Computer Science and Engineering

Faculty of Engineering

The University of New South Wales

Few Labels Classification in Text - Exploring Machine Learning Models for Cardiogenomic Diagnosis

by

Niranjana Arun Menon

Thesis submitted as a requirement for the degree of

Bachelor of Engineering in Computer Engineering

Submitted: November 2025

Supervisor: A/Prof. Imran Razzak Student ID: z5417727

Abstract

Cardiovascular disease (CVD) is defined by its complex interactions between multiple factors involving genetic, environmental, and lifestyle preferences of an individual. Genomic profiling can be paired with electrocardiography to identify markers associated with the risk of developing the CVD . However, it is often difficult to translate this information from accurate diagnostic tools, particularly in contexts where annotated data is limited.

This thesis primarily focuses on investigating machine learning-based models for classification of cardiovascular diseases (CVDs) using genetic markers and electrocardiogram data, focusing on both traditional classifiers and transformer-based approaches. By leveraging these structured representations from high-dimensional yet sparsely labeled single nucleotide polymorphism (SNP) data, we explore how various models can learn meaningful patterns in low-label environments. By defining cardiac classification as a low-resource learning task, it enables us to evaluate across sets of diverse phenotypic profiles, offering insights into the strengths and limitations of different algorithmic strategies. The study also incorporates temporal non-confirmed clinical diagnosis through the construction of patient-level knowledge graphs, enabling multi-modal integration and longitudinal modeling. The ultimate goal of the thesis is to support clinical early diagnosis, even in the absence of abundant annotated clinical data.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Professor Imran Razzak, for his unwavering guidance, encouragement, and mentorship throughout this project. His support has been detrimental in shaping the direction and quality of my research. I am also sincerely thankful to Professor Sonit Singh for his insightful feedback and continued support, which I will always value.

My appreciation extends to the three PhD students I had the privilege to work with as a part of this thesis as well. Their knowledge, patience, and collaboration have been vital in helping me grasp the technical aspects of my work.

I am equally indebted to the wider academic community, whose scholarly contributions have formed the foundation of this research. The dedication of these researchers has enabled me to build upon a rich body of literature in pursuit of my thesis goals.

Finally, I would like to thank my family for their unconditional support and encouragement, and for always believing in me.

Abbreviations

BE Bachelor of Engineering

PhD Doctorate of Philosophy

CVD Cardiovascular Disease

SNP Single Nucleotide Polymorphisms

rsID Rapid Stain Identification

ECG Electrocardiogram

GWAS Genome Wide Association Studies

NHS National Health Service

WHO World Health Organisation

HPP Human Phenotype Project

ML Machine Learning

LoRA Lower Rank Adaptation

PEFT Parameter Efficient Fine-Tuning

LLM Large Language Models

SVM Support Vector Machines

AUC Area Under the Curve

MCC Matthew's Correlation Coefficient

PCA Primary Component Analysis

KG Knowledge Graph

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Motivation

In recent years, progress in machine learning has opened new possibilities for interpreting complex biological data, particularly for genetic risk stratification. Cardiovascular diseases (CVDs), the leading cause of death globally [WHO25, NHS17], have strong but complex genetic components that influence an individual's susceptibility to CVDs. However, these genetic components are not yet fully understood owing to its complex nature.

Similarly, there has been influx of availability in high-dimensional genetic datasets consisting of Single Nucleotide Polymorphisms (SNPs) in the recent years, including the HPP dataset [Phe21]. This presents for an exciting opportunity to explore complex genotype-phenotype relationships. While large-scale genomic studies such as GWAS [EBI] consist of numerous genetic variants associated with CVDs, using this information for accurate disease classification still poses as a challenge. This is especially apparent in scenarios where confirmed diagnosis can be scarce or incomplete. The limited availability of labeled and clinically confirmed outcomes can become a significant challenge for implementing traditional supervised learning techniques that depend on large volumes of annotated data. This limitation highlights a critical research gap,

which we aim to address in this thesis.

1.2 Thesis Goal

The aim of this thesis is to develop a multimodal, few-label pipeline for CVD classification, integrating structured genomic and ECG data with machine learning and LLM-based reasoning to improve predictive accuracy and interpretability.

1.3 Key Objectives

The key objectives we aim to address are as follows:

1. Develop a multimodal cardiac disease classification pipeline integrating SNPs, ECG data, biomarkers, and metadata.
2. Benchmark traditional ML models against LLMs for few-label classification tasks.
3. Explore tier-based learning strategies for under-labeled datasets.
4. Implement Chain-of-Thought reasoning within LLMs for biomedical inference.
5. Construct knowledge graphs to augment predictions and improve interpretability.
6. Evaluate temporal cardiac risk prediction across multiple patient visits.

We aim to first introduce key biomedical concepts and relevant literature to establish the foundation for our dataset and methodology. Overall, the research aims to improve automated methods that bridge genotype data with clinical phenotypes, aiding clinical identification of individuals at risk based on their genetic profiles.

Chapter 2

Background

Note. Parts of this chapter were originally completed in Thesis A. The sections on biomarkers are newly developed, as the approach was proposed in the Thesis A Report.

To achieve the goals mentioned in Chapter 1, it is important to define three key concepts in the field of cardiogenomics. This chapter introduces the two primary data modalities utilized throughout this research: Electrocardiograms (ECGs), Single Nucleotide Polymorphisms (SNPs) and Biomarkers. These modalities will then be incorporated with multimodal modeling, machine learning and LLM based approaches, more of which will be discussed in Chapter 3.

2.1 Electrocardiogram Data

Electrocardiograms (ECGs) measurements are waveforms used as a non-invasive diagnostic tool to measure heart rates over time. Abnormalities in these waveform features are primarly used to detect various cardiovascular conditions such as myocardial infarction (MI), arrhythmias, and hypertrophy.

A typical ECG waveform, as illustrated in Figure 2.1, consists of several key components that reflect the heart's electrical activity during each cardiac cycle.

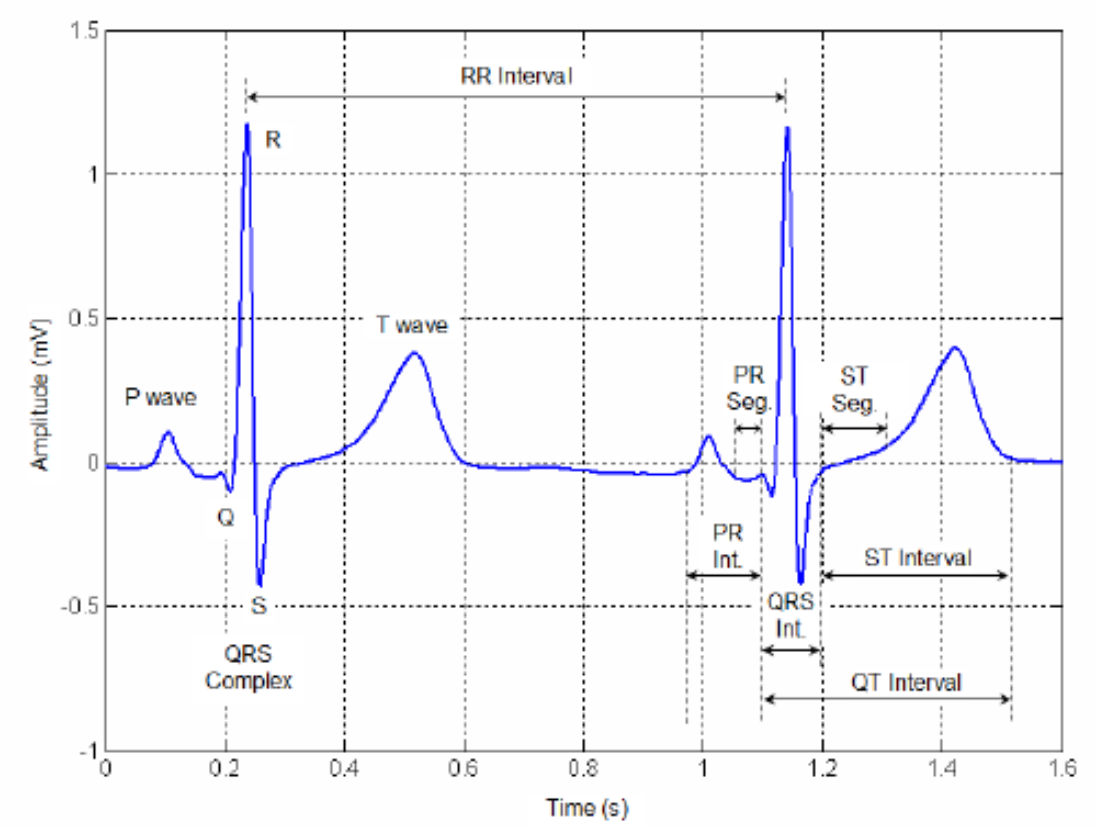

Figure 2.1: Sample ECG waveform obtained from [MBLA12].

Here, the features are critical for diagnosing various cardiac abnormalities are as follows, defined by Maggio et al. [MBLA12]:

- **P wave:** It represents the atrial depolarization and indicates the start of electrical activity in the heart.
- **QRS complex:** It represents the ventricular depolarization and is used for assessing conduction and rhythm abnormalities.
- **T wave:** It represents the ventricular repolarization and is important for detecting electrolyte imbalances.
- **ST segment:** It represents the period between ventricular depolarization and repolarization (QRS and T wave); an elevation or depression during this segment could signal a myocardial infarction.

- **QT interval:** It measures the duration from the start of ventricular depolarization to the end of repolarization; a prolonged QT interval can mean a heightened risk of arrhythmias.
- **R-R interval:** It represents the time between successive R-waves, which is then used to calculate heart rate and identify rhythm disturbances.

The patterns of the ECG readings can be used for effective CVD risk identification. However, ECG data alone may be insufficient for a comprehensive reading, especially in cases of phenotypic overlap such as Atrial Fibrillation and Hypertrophic Cardiomyopathy [LAH+18]. For this reason, we consider integrating ECG signals with genomic data to enhance diagnostic precision.

2.2 Genetic Markers: Single Nucleotide Polymorphisms (SNPs)

Single Nucleotide Polymorphisms (SNPs) are known as single base substitutions in the DNA sequence which occur at specific genomic loci and represents the most common form of genetic variation in the human genome [GLG+]. Each SNP is identified by a unique reference SNP cluster ID called Rapid Stain Identification (rsID), which helps in tracking and interpretating diseases that are associated with these SNPs. The extensive information about each individual SNP can be found through multiple public databases such as dbSNP [NCB98] and GWAS catalogs [GWA08]. SNPs are crucial markers for CVDs because they influence gene regulation, protein function, and cellular pathways, all of which can affect the blood regulation to heart and cause CVDs.

Conventional methods such as polygenic risk scores (PRS) and classical machine learning models are often limited in their ability to capture complex, non-linear interactions among genetic variants. These approaches typically rely on additive effects of the variants and may fail to account for higher-order feature dependencies. In contrast, recent developments in deep learning, particularly transformer-based architectures, have

demonstrated strong potential for modeling high-dimensional genetic data. Authors such as Mieth et al. [MKR+16] suggest that these transformer based models can leverage contextual and regulatory features to enhance prediction accuracy.

We also aim to explore how such deep learning models, including transformer-based architectures, helps to uncover latent patterns within SNP data relevant to cardiovascular risk, seeking to improve the classification of cardiac diseases in settings where annotated data is sparse. But first, it would be helpful to realize how SNP variants and ECG phenotypes correlate for CVD risk prediction.

2.3 Modeling SNP Variants and ECG Phenotypes for CVD Risk Prediction

Cardiogenomics data can be modeled using genome-wide association studies (GWAS), which have consistently identified genetic loci that contribute to cardiovascular disease (CVD) risk. Figure 2.2 illustrates the interrelations between genetic variants and ECG features; an intersection highly relevant to understand to achieve the goals of this thesis.

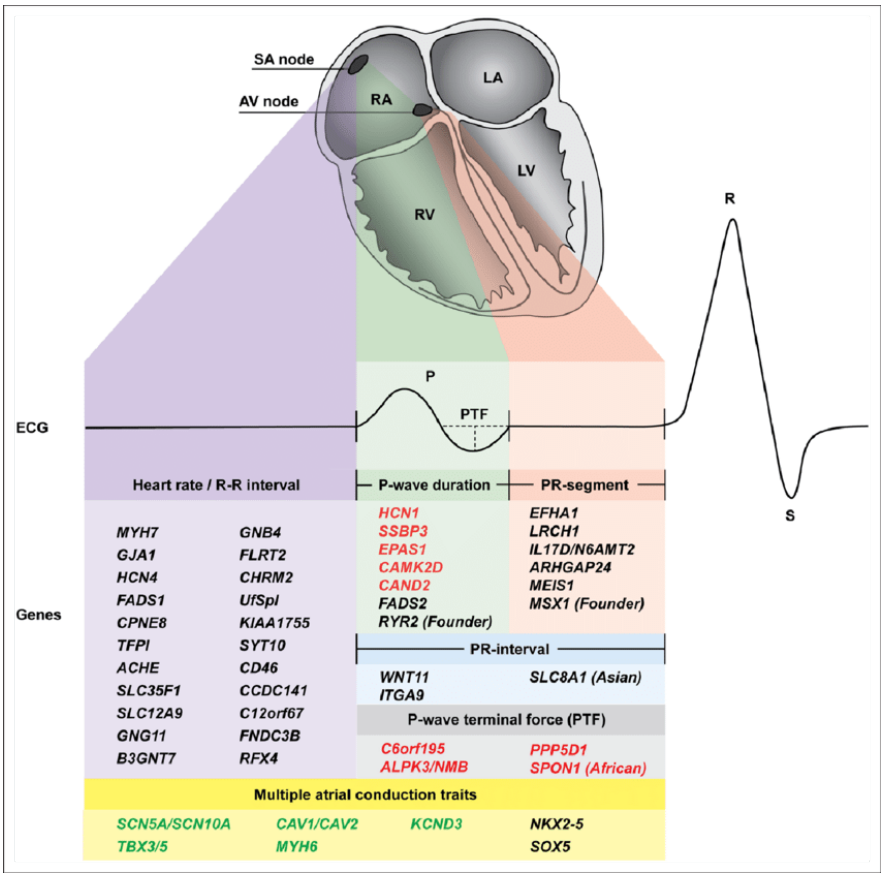

Figure 2.2: Illustration of associations between ECG parameters and genetic loci [CMY+17].

As introduced in Section 2.1, ECG phenotypes such as R-R interval irregularities play a critical role in diagnosing arrhythmias such as atrial fibrillation (AF). In the context of the Figure above, the SNPs influencing atrial conduction and rhythm regulation gain importance. For example, the gene MYH7, shown in Figure 2.2, influences the R-R interval, very inline with influencing behaviour of AF (that relies on irregular R-R intervals as one of the ECG markings), as demonstrated by Lee et al. [LAH+18]. Although MYH7 is traditionally known for its role in hypertrophic cardiomyopathy, Lee et al. provide evidence which suggests that its variants may contribute to Atrial Fibrillation as well.

However, MYH7 is not the sole variant of interest for Atrial Fibrillation. Over the years, multiple loci and SNPs have been identified across studies , and comprehen-

sive databases such as dbSNP [NCB98] and GWAS Catalog [GWA08][1] provide access to these variants associated with both monogenic and polygenic forms of CVD (also mentioned in Section 2.1).

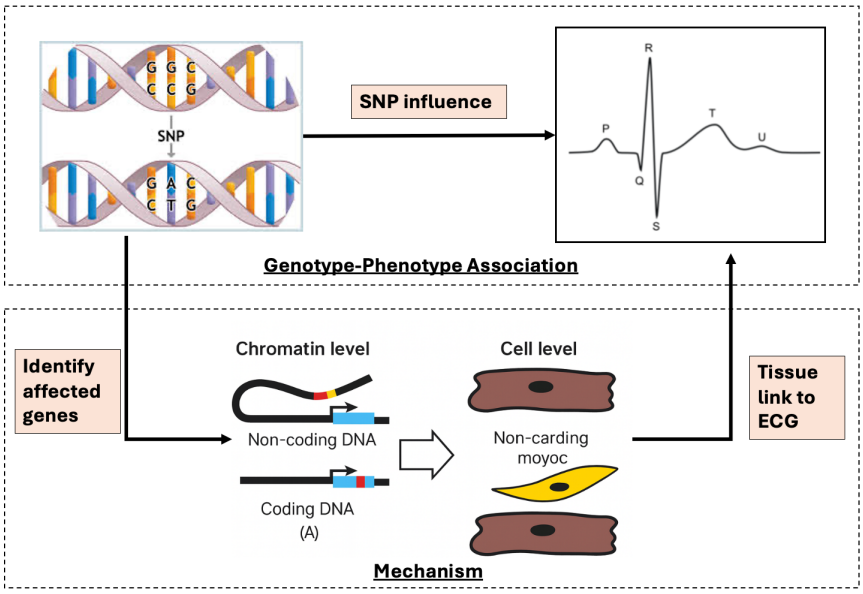

Figure 2.3: Linking SNP Variants and ECG Phenotypes for Explainable CVD Prediction.

Figure 2.3 also illustrates the relationship between cardiac gene expression and electrophysiological phenotypes derived from electrocardiogram (ECG) signals, where the chromatin level DNA influence cell level tissues, which in turn influences the behavior of the ECG signals.

2.4 Biomarkers

Biomarkers provide quantifiable biochemical signatures that reflect the metabolic, vascular, and physiological state of an individual. These can be quite crucial to uncover traits for cardiovascular diseases. In cardiovascular disease (CVD) research, biomarkers serve as early indicators of risk, offering insight into lipid metabolism, glucose regula-

[1]While these databases may have been founded in 1998 and 2008 respectively, they are updated with information as of current date and year.

tion, inflammation, and overall cardiometabolic balance [NHW+25]. Unlike diagnostic labels, which often capture late or advanced stages of disease, biomarkers enable the detection of subtle deviations from metabolic homeostasis, thereby supporting preventative and personalised risk stratification. This can be quite useful when modelled together with ecg phenotypes and genetic information.

2.4.1 Lipid Biomarkers

Lipid-related biomarkers such as triglycerides and high-density lipoprotein (HDL) are foundational to assessing cardiovascular health. Elevated triglyceride levels are associated with atherogenic dyslipidaemia and metabolic syndrome, increasing the likelihood of coronary artery disease [LT19]. HDL cholesterol is considered cardioprotective due to its role in reverse cholesterol transport as per Erol and Gulec[GE20]. Together, these lipid markers can form a biochemical profile that solely reflects on an individual's susceptibility towards vascular imbalance and long-term cardiac risk.

2.4.2 Glucose and Metabolic Biomarkers

Another critical component of cardiovascular profiling is markers related to glucose metabolism, including fasting glucose and insulin sensitivity. Even mild elevations in fasting glucose have been shown to correlate with increased CVD incidence, which reinforces the sheer importance of metabolic biomarkers as early indicators of cardiometabolic stress, as stressed by Poznyak et. al [PLP+22].

2.4.3 Dietary and Nutritional Biomarkers

Diet also plays a crucial role in shaping cardiometabolic biomarkers. Nutrient intake, which includes macronutrients such as carbohydrates, fats, and proteins, directly influences lipid profiles, inflammatory pathways, and glucose regulation [ESGA25]. Nutritional biomarkers can therefore provide valuable contextual information about

lifestyle and metabolic load. For example, high dietary fat consumption may elevate triglycerides, while excessive sugar intake can impair glucose tolerance. Incorporating diet-based data into cardiovascular models enables a more holistic representation of individual health, capturing behavioural factors that interact with biological risk markers.

In summary, integrating ECG phenotypes with SNP-based genomic data holds a significant potential to advance precision CVD classification. While ECGs offer a snapshot of electrical activity and cardiac function per cardiac cycle, SNPs provide rich insights into the underlying heritable risk. Combining these two modalities together through knowledge graphs and transformer-based architectures, while enhanced with biomaerks, can drastically improve classification, risk prediction, and individual-based report generation for CVDs influenced by genetics.

Chapter 3

Literature Review

Note. Similar to Section 2, parts of this chapter were originally completed in Thesis A. The section on semi-supervision is newly developed, as the approach was slightly modified during Thesis B.

3.1 Summary

This chapter provides a detailed literature review of the core techniques, models, and theoretical frameworks that will be relevant to this thesis.

The review will focus on four key areas that are relevant to the proposed pipeline: knowledge graphs, machine learning classifiers, large language models (LLMs), semi-supervised clustering and time series analysis. When modeled together or evaluated against each other (in the case of the classifiers and LLMs), they help strengthen the interpretability of the multimodal cardiogenomic datasets.

3.2 Knowledge Graphs

A knowledge graph approach was considered over dictionary mapping or key-value based approaches to ensure that complex mapping of diseases by their associated factors is preserved. This approach will allow machine learning methods to better structure and interpret relational information between diseases, genetic variants, and other risk factors, which will then support predictions tailored to each participant.
Previously, classical approaches such as BERT-based models were considered to be part of the project pipeline as well. However, BERT based models typically require extensive labeled training data, as mentioned by studies lead by Wu et al. [WX25] and Rietberg et al. [RNG+23], which cannot be applied in our case since we established in Chapter 1 that our focus follows datasets with limited labels for disease-specific tasks.

However, recent advancements in this area has allowed us to discover a novel pipeline to address these limitations. One such study, conducted by Xu et al. [XGX+24], constructs a process involves schema design, information extraction, and knowledge completion, each enhanced through carefully crafted task-specific prompt templates and the TwoStepChat method. This TwoStepChat method sidelines every other BERT model for knowledge graph construction, as mentioned in the study.

To elevate the efficiency of the method, the study provides examples of prompts for both the Vanilla (which are simple few-shot prompts) and TwoStepChat approaches, summarized as follows:

Vanilla Prompt The Vanilla Prompt method involves using a large language model with a single, broad instruction to extract both entities and their relationships from unstructured clinical text. An example of a vanilla prompt could be:

> "Extract the medical entities and relationships from the following passage."

This approach was said to struggle with complex medical texts and performs poorly in recognizing unseen entities or relations, particularly when the input data is diverse or

deviates from the training set.

TwoStepChat However, in TwoStepChat, the extraction process is divided into two sequential tasks:

1. **Entity Recognition Phase:** The model is first instructed to identify and categorize all relevant medical entities in the text. For example:

   "Identify all medical entities (e.g., diseases, symptoms, biomarkers) mentioned in the following passage."

2. **Relation Extraction Phase:** Then, the model is prompted to extract relationships between the previously identified entities. For example:

   "Given the identified entities, extract and describe the relationships between them."

Experimental results presented as a part of the study focuses on evaluating all the BERT models for Vanilla Prompting against the TwoStepChat model (powered by Chat GPT 3.5 as the chosen LLM for extraction and description purposes) and illustrates how TwoStepChat outperforms both vanilla prompting and fine-tuned BERT models. Additionally, it is said to also reduced manual annotation requirements by approximately 65%, offering an optimized solution for building clinical knowledge graphs in data-limited settings.

Given these findings, we aim to integrate a similar methodology and construct knowledge graphs that incorporate genetic, clinical, and phenotypic data per participant, and map them based on the relationships between participants, which will help enhance interpretability and predictive modelling.

3.3 Traditional Classifiers for Cardiovascular Disease Prediction

Following the construction of knowledge graphs is the choice of models for classification purposes. In this thesis, we aim to evaluate traditional machine learning classifiers to understand their effectiveness in cardiovascular disease (CVD) risk prediction.

Latimer et al. [SA24] critically assess the limitations of standard CVD risk prediction models, noting that methods like QRISK3 and the Framingham Risk Score underperform in data populated diverse patient information.

To improve on such methods, the authors choose to incorporate several machine learning models such as logistic regression (LR), random forests (RF), support vector machines (SVM), and deep neural networks (DNNs) instead of the standard prediction models, all trained on large-scale UK Biobank data. They reported improved predictive performance on these models over the standard ones and emphasize the importance of model calibration, subgroup analysis, and external validation.

While complex models like DNNs showed marginal performance gains over simpler classifiers, the study highlights key challenges such as data imbalance and model interpretability can pose an issue in terms of classification accuracy. By observing this insight, we lead to the decision of deciding not to employ standalone neural networks for classification, and instead focus on interpretable and generalizable alternatives.

Another instance of classification analysis is also presented by Gharebakhshi and Fathian [KS17], as they address the challenge of early diagnosis of coronary artery disease (CAD), particularly given its asymptomatic onset. They propose a model that combines Random Forests with Genetic Algorithms for feature selection, achieving a validation accuracy of 93.2%. This hybrid approach outperformed standalone SVM models, demonstrating the strength of combining ensemble methods with optimized feature sets. Notably, their model achieved high stability across multiple datasets while reducing feature dimensionality, which is a key advantage for interpretability and clinical applicability.

However, in a more recent study, Peng et al. [PSZ22] investigate gene expression profiling for CAD diagnosis using SVM, RF, and LR. Using weighted gene co-expression network analysis (WGCNA) and recursive feature elimination (RFE), they identify optimal feature genes (OFGs) for classifier training. Among these models, SVM was said to perform best, achieving an internal validation AUC of 0.996 and external validation AUC of 0.813. Although RF and LR showed slightly lower performance, they still yielded competitive results. We find that the study highlights the promise of integrating advanced feature selection techniques with machine learning for genetic-based diagnosis, while acknowledging the absence of clinical data integration and the need for further biomarker validation.

Overall, the genetic classification models for CAD is nuanced. Even though SVM models can achieve high internal AUCs, Random Forests incorporated with feature selection strategies can demonstrate better external generalizability. Based on these findings, we will consider incorporating both SVM and RF classifiers as options (with and without feature optimization) and evaluate their respective strengths in predicting cardiovascular disease outcomes.

3.4 LLMs in CVD

Apart from traditional classifiers, large language models (LLMs) have an incredible potential for CVD classification tasks due to its powerful computational capabilities and flexible reasoning. In the more recent years, LLMs have been explored across a wide spectrum of biomedical applications, from clinical decision support to biological sequence analysis. However, unlike conventional models that depend heavily on hand-engineered features or domain-specific preprocessing, LLMs can infer complex relationships directly from raw or minimally processed data, making them especially well-suited for high-dimensional biomedical datasets without prior domain-specific preprocessing.

This section aims to highlight the reasons why LLMs are computationally useful in the

context of cardiovascular disease (CVD) risk assessment, with a particular emphasis on genomics and clinical records to aid our research.

3.4.1 LLMs for Electronic Health Record (EHR) Analysis

There is a huge potential for LLMs in extracting clinically relevant information from free-text electronic health records (EHRs). In a study conducted by Gu et al., [GSL+25], the authors utilize LLMs to automate information extraction across rich clinical datasets without the need for extensive fine-tuning. Their approach primarily relies on zero-shot and few-shot prompting, enabling LLMs to accurately identify clinical concepts, temporal relationships, and patient outcomes from heterogeneous EHR corpora.The LLMs utilized were general purpose ones, such as GPT-3.5 and GPT-4 and were compared against traditional rule-based and supervised systems across various information extraction tasks.

Notably, LLMs achieved competitive or superior performance in identifying diagnoses, treatments, and timelines, even in scenarios lacking task-specific training data. One of the key strengths lie on its scale to incorporate adaptable prompts and reduce the need for extensive annotation or retraining when applied to new clinical domains or healthcare systems.

However, it highlights the limitations of LLMs, particularly surrounding hallucination risks and sensitivity to prompt phrasing. Mitigation strategies such as refined prompt engineering and post-processing rule are proposed to enhance model reliability, something that we intend to utilize in our proposed plan.

3.4.2 LLMs in Genomics

Similarly, LLMs have shown incredible potential in identifying genetic patterns, predicting protein structures, and interpreting biological responses. A recent study conducted by Sarumi et al. [SH24] aimed to investigate whether general-purpose LLMs can pre-

dict protein behavior under radiation exposure and disease conditions using zero-shot and few-shot prompting techniques. The LLMs here were presented with contextual descriptions of proteins, diseases, and biological stressors such as radiation, and tasked with predicting cellular responses such as up regulation or down regulation. The results suggest that GPT-3.5 and similar models could extract biologically relevant patterns with surprising accuracy, even without fine-tuning; with the performance at its highest prompts were engineered with a structure specifying the condition, biological pathway, and expected response.

These insights thus presents the ability of LLMs to generalize across unseen protein-disease-stimulus combinations. Even though it cannot replace a mechanistic models, the zero-shot framework can provide for a scalable method for hypothesizing protein responses in contexts where experimental data are sparse or unavailable, something we can also utilize in our project plan.

3.4.3 LLMs for CVD Risk Assessment

Exploring LLMs in both EHR analysis and genomics naturally leads to the focus of LLMs in relation to cardiovascular disease (CVD) risk assessment. Studies conducted by authors such as Karim et al. [KMIA24] demonstrates the positive effect of LLMs for improved disease classification accuracy across multiple domains, including CVD. This is done by employing multiple fine-tuned transformer-based models to classify patient records into major disease categories.

The models demonstrated significant improvements in comparison to traditional baselines by capturing subtle clinical features relevant to each disease class. This is particularly relevant for cardiovascular conditions, where early detection depends on the ability of the LLM to interpret complex, multimodal signals from multiple data sources.

In conclusion, the results from all three research articles all agree on the abilities of LLMs to capture nuanced patterns over multi disease prediction tasks and reinforce their value in risk assessment pipelines for CVD. This positive reception allows us to

incorporate LLMs as another major option for CVD classification purposes, all the while trying to mitigate any hallucination effects that could occur at the time of evaluation.

3.5 Semantic Aware Pseudo-Labelling

Semi-supervised learning (SSL) has become an essential standard for leveraging large quantities of unlabeled data when annotated corpora are scarce. Pseudo-labeling and self-training remains two of the most widely adopted SSL strategies in natural language processing (NLP). Classical self-training operates by iteratively training a model on labeled data, assigning labels to unlabeled instances, and retraining on the newly labeled data [Yar95]. While conceptually it is quite simple, this approach is highly sensitive to error accumulation. This means that early misclassifications can reinforce themselves, especially when the model encounters ambiguous or low-density decision regions.

To overcome these limitations, recent literature has introduced more structured pseudo-labeling frameworks. One significant advancement is presented by Lee et al., who incorporate confidence-based filtering, where models only accept pseudo-labels exceeding a predefined certainty threshold [Lee13]. However, this threshold-based methods can still suffer from two critical issues: (1) boundary ambiguity, where samples near decision boundaries receive unreliable pseudo-labels, and (2) class imbalance amplification, where majority classes dominate pseudo-label assignments and skew the training distribution.

Yang et al., directly address these challenges by proposing the ProtoS^2 framework, which is a prototype-guided pseudo-labeling framework [WY23]. In their methodology, they introduce two key components: Prototypical Cluster Separation (PCS) and Prototypical-Center Data Selection (CDS). PCS constructs class prototypes from labeled instances and then uses a prototype-anchored contrastive objective to pull representations toward their class prototype while pushing them apart from others. This results in more compact and well-separated class clusters, reducing the likelihood of assigning pseudo-labels to ambiguous boundary points. CDS further refines pseudo-

label reliability by selecting only those unlabeled examples that lie close to their corresponding class prototype, thereby enforcing distribution-aware pseudo-label filtering and mitigating class imbalance effects.

The empirical results from Yang et al. suggest substantial improvements across multiple semi-supervised text classification benchmarks as their methods outperfom baseline methods by 6.8%-13.8%, thus highlighting the importance of structured representation learning in pseudo-labeling pipelines. These approaches collectively illustrate that effective pseudo-labeling requires not only confidence estimation but also principled modeling of representation space geometry and class distribution.

Overall, the evolution from classical self-training to prototype-guided pseudo-labeling reflects a shift toward more reliable, structure-aware semi-supervised methodologies. These advances are particularly relevant for domains with limited annotated data, where reducing error propagation and class bias is critical for stable model performance, and was thus considered in our use-case to balance the few-labels environment and provide the unlabelled instances with meaningful labels.

3.6 Time Series Based Analysis

With the models successfully selected, we then consider the power of time-series investigation for CVD classification. In particular, we believe it to elevate early diagnosis in patients through forecasting based on the patient data provided and the history of the patients involved. Through extensive research, Artificial Neural Networks were found to be incredibly powerful to collate and compute such information.[1]

Artificial neural networks (ANNs), particularly feedforward multilayer perceptrons (MLPs), are shown to have strong potential for modeling and forecasting physiological signals through prior work, which highlights that predicting ECG waveforms can

[1]It should be noted that while standalone DNNs were deemed unsuitable for classification due to interpretability concerns in the previous section, shallow ANN architectures should be considered for time series modeling due to positive evidence.

aid clinicians in early intervention. For example, Kwembe et al. [BMBP19] argue the importance of ECG prediction to assist specialists in providing early measures against CVD and allow the demonstration of a feed-forward neural network that can effectively predict ECG signals. Prakarsha and Sharma further build on this proposed method [RPS22] and suggest that the application of ANN to forecast ECG time series will result in a high accuracy, thus aiming to improve upon traditional adaptive filters.

Some of the key takeaways from the paper include:

- The perceptron network in ANN is trained on past ECG samples to predict future signal values, which allows leveraging the approximation capability of MLPs for nonlinear time series.
- The ANN was first trained to forecast the waveform several steps ahead, effectively denoising as it predicted.
- Forecast accuracy (as percentage agreement or correlation with true waveform) and signal-to-noise ratio (SNR) of the predicted signal were the main metrics used to predict the efficiency of ANN's performance. The performance was compared against a standard Least-Mean-Squares (LMS) adaptive filter baseline which are traditionally used for time-series analysis.

It was noted that the ANN massively outperformed a traditional LMS filter. The ANN achieved 95.72% accuracy versus 79.0% for LMS, and produced a predicted signal with SNR$\approx$ 29.7 dB (versus 16.3 dB for LMS). These results indicate that neural network forecasts match the true ECG waveform and also effectively reduce noise, yielding cleaner signals than the traditional filter which can be particularly useful for precisely forecasting ECG trends that may help clinicians identify arrhythmias in advance.
This observation allows us to consider incoroporating ECG prediction using neural networks, which will allow enhancement of early diagnosis or risk assessment of cardiovascular conditions.

In summary, we have identified key technologies relevant to multimodal CVD risk prediction. By utilizing the proposed TwoStepChat Approach, knowledge graphs can be

used to structure patient-level relationships by utilizing genomic, clinical, and phenotypic factors. Meanwhile, LLMs and traditional classifiers will be explored and compared against each other for scalable information extraction and classification. Semi-Supervised clustering will help balance the few-labeled dataset into meaningful clusters, and finally, the time series methods will enhance the modeling of dynamic ECG signals. Together, these components form the methodological backbone of this thesis, detailed further in the next chapters.

Chapter 4

Datasets

This chapter describes the multimodal datasets used throughout this thesis, including genetic variants (SNPs), electrocardiogram (ECG)–derived phenotypes, clinical biomarkers, nutritional information, and tier-based diagnostic labels. The testing is conducted on the Pheno.AI Human Phenotype Project (HPP) [Phe21] dataset, which consists of a rich longitudinal cohort inherently designed for genomics–phenotype research.

4.1 Overview of the Pheno.AI Dataset

This PhenoAI repository consists of cohort data from individuals residing in the Asia/-Jerusalem time zone, meaning that the population is not geographically stratified across multiple time zones. As such, temporal or environmental confounding factors linked to circadian rhythm variation across regions are minimized. This unified multimodal dataset provides a foundation for jointly modeling genetic variation and electrophysiological signatures, which facilitates more effective cardiovascular risk stratification. Participants also have information about multiple visitations, which allows us to also pursue temporal analyses of cardiac risk. Out of the total thirty-four datasets present, the following datasets were used in our use-case: `005:diet-logging`, `007:blood-pressure`,

`014:human-genetics`, `015:ecg` and `021:medical-conditions`. [1]

4.2 Genetic Variant Data (SNPs)

In the `015:human-genetics` dataset, the genotype data is stored in PLINK binary format (`.bed`/`.bim`/`.fam`) under each chromosome file, and there are 22 chromosome files present in total. The relevant features include unique rsIDs, and each rsId `.bed` matrix is encoded as the following under a participant x SNP matrix, as shown in the Table 4.1. There are 33 million SNPs in total.

Table 4.1: Example of participant genotype and phenotype data in a .bed matrix.

FID	**IID**	**PAT**	**MAT**	**SEX**	**PHENOTYPE**	**rs1**	**rs2**	**rs3**	**rs4**	**rs5**
ID1	ID1	0	0	1	1	0	1	2	1	0
ID2	ID2	0	0	2	1	2	2	0	1	1
ID3	ID3	0	0	1	2	1	0	1	2	2

Note: Allele coding is as follows: where 0: homozygous major allele ("term absent"); 1: heterozygous ("term present once"); 2: homozygous minor allele ("term present twice").

As a part of visualizing the split of the rich genetic data present, we also showcase the modeling of MAF distribution in figure 4.1 across 22 chromosomes. For context, the frequency at which the second most occuring allele in an SNP appears in a population is known as minor allele frequency (MAF) [maf], and MAF allows us to determine which SNPs are common or uncommon. As observed from this diagram, we can see that there are quite a number of variants around the MAF values of $0.0 - 0.1$ as it forms as the majority of the distribution, which means they are rare variants (as they are less frequent than most variants).

These variants are interesting because they can have larger biological effect sizes, but also introduce higher noise, especially when dealing with smaller sample sizes. By setting a MAF cutoff at 0.01, we basically strike a neat balance, as we keep informative

[1] Although a family-linkage dataset existed and could have supported analyses of inherited CVD risk, it predominantly contained room-mate information rather than familial relationships at the time of analysis, so it was not used.

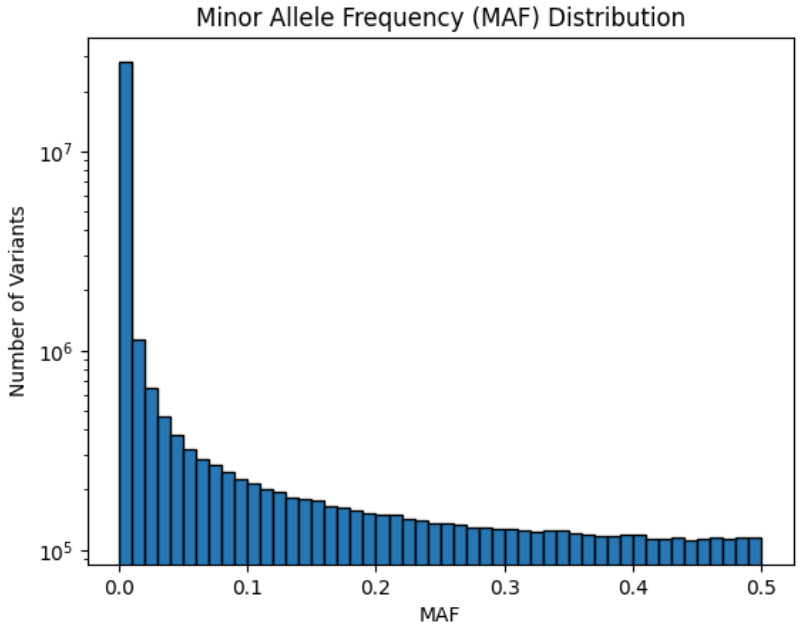

Figure 4.1: MAF distribution prior to genetic QC.

variants while minimizing potential noise.

In addition to applying MAF filters, we also implemented Hardy–Weinberg Equilibrium (HWE) and Genotype Posterior (GP) filters [ABT20], which have been previously used for genetic-based analysis. These filtering steps left us with 9.8 Million SNPs for our experiments, substantially reduced the number of variants available and allowing us to retain the rarer variants that could have a huge impact on a disease. The cohort distribution before and after QC application is summarized by Table 4.2.

Table 4.2: Cohort distribution and SNP/ECG feature counts before and after quality control.

Group	**Participants**	**SNP variants (raw)**	**SNP variants (post-QC, approx.)**	**ECG features (QC)**
Confirmed cardiac labels	350		≈ 387,000	12
Some cardiac connection	816	33,000,000 (dataset-level)	≈ 554,000	12
Possesses no prior cardiac condition	9426		≈ 6,659,000	12
Total	10160	33,000,000	≈ 9,800,000	12

4.3 Electrocardiogram (ECG) Phenotype Data

The `014:ecg` dataset in Pheno.AI consists of consists of time-series data from standard 12-lead ECGs collected at each participant visitation.

Figure 4.2 depicts the sample phenotypic distributions of ECG data across the participant pool around the ages 40-60. This is very in line with previous findings that suggest that people over the age of 50 are very likely to develop a CVD conducted by NHS **??**, which makes it a very relevant dataset to observe.

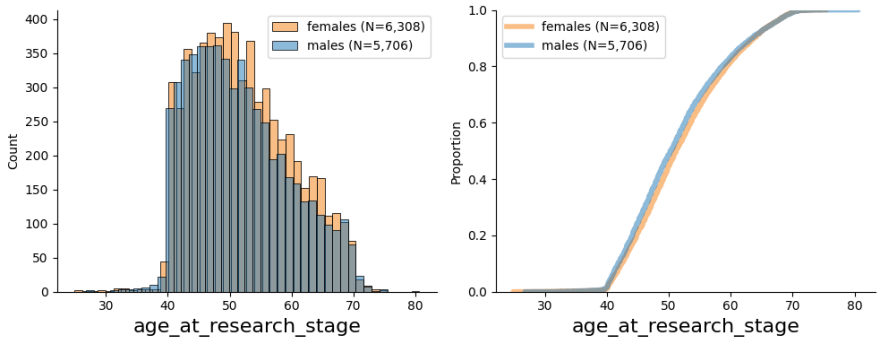

Figure 4.2: Example ECG feature distributions.

4.4 Medical Conditions Data

The `021:medical-conditions` dataset actually serves as the clinically-annotated dataset and provides labels for identifying cohorts with cardiovascular diseases.

Figure **??** represents a fraction of the split of participants with known cardiac conditions. As seen from the image, we can infer that not many people possess CVDs, which technically puts the PhenoAI dataset under the few-labels environment and justifies our findings accordingly. Thus, this makes it a very relevant dataset to observe as well.

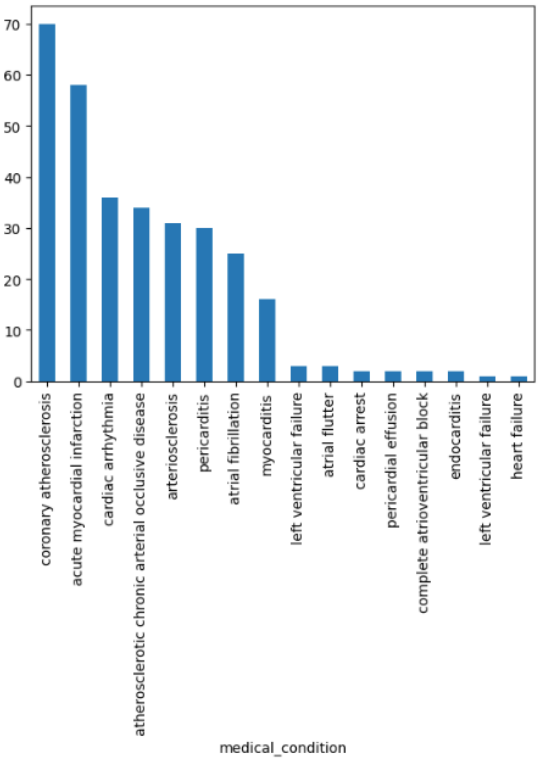

Figure 4.3: Cohort of Participants with Cardiovascular Conditions

4.5 Biomarker Distribution

Similar to the previous datasets, the biomarker and nutritional datasets `005:diet-logging` and `007:blood-pressure` also have distributions that exhibit meaningful variability between sexes and across time, while also showcasing variability of macro nutrients as shown in Figures 4.4 and 4.5. This also makes them relevant datasets for the usage of our evaluations.

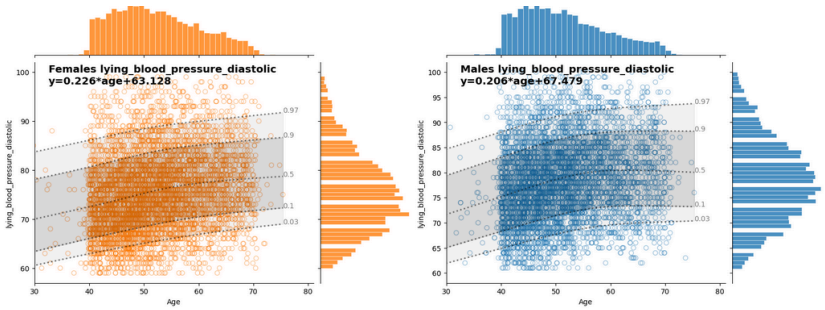

Figure 4.4: Blood test distributions (sex-stratified).

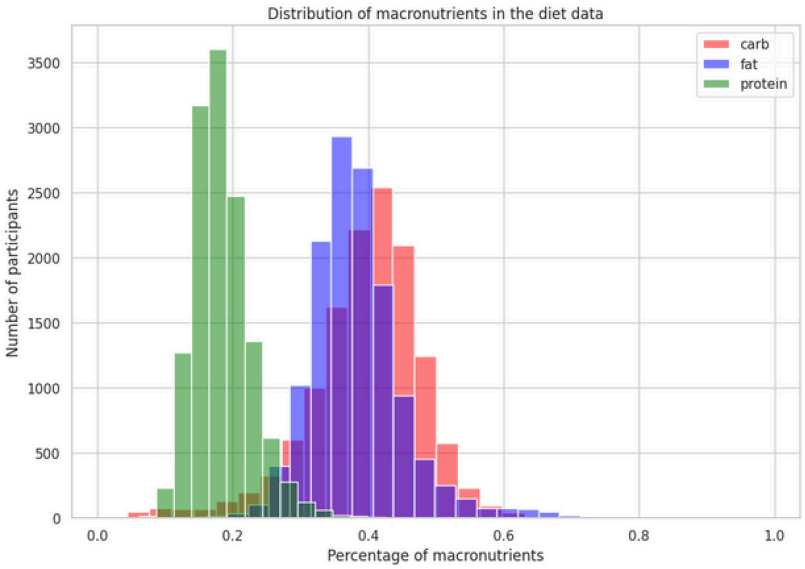

Figure 4.5: Macronutrient distribution across the cohort.

4.6 Participant Demographics and Longitudinal Structure

In total, there are 10,160 participants present in this dataset, and their information is split across years. However, not every participant has multi-year visitation reports. In fact, only a handful of participants consists of three year information, much less two. Table 4.3 summarizes this split, while table 4.4 presents the number of visitations available for each individual, ranging from one to three annual measurements. As we can see, there are only 352 participants that consists of three-year visitation information, while only 2312 participants have two-year visitation information. While it only a fraction of the full cohort, this information can be valuable for risk prediction, as we can use the cohort with three year annual data to forecast potential risks for cohorts who may have similar characteristics to these participants but only have a one year or two year visitation record.

Table 4.3: Distribution of participants across Tiers and visitation counts

Visitation	**Tier 1**	**Tier 2**	**Tier 3**	**Total**
0	350	2652	7128	10160
1	80	693	3042	3815
2	31	138	1036	1205

Number of Visitations	**Participants**
3 Visitations	352
2 Visitations	2312

Table 4.4: Participant distribution across longitudinal visitations.

4.7 Summary

In summary, the Pheno.AI dataset provides a rich multimodal foundation for cardiogenomic modelling. However, its high-dimensional, sparsely labelled, and longitudinal nature presents significant challenges, motivating the use of traditional ML, LLMs, and knowledge-graphs in subsequent chapters.

Chapter 5

Methodology

This chapter outlines the overall pipeline used to develop and evaluate a multimodal, few-label system for cardiovascular disease (CVD) risk prediction. Our approach combines SNP genotypes, ECG-derived phenotypes, and biochemical biomarkers with traditional machine learning classifiers, modern large language models (LLMs), semi-supervised learning, and clinical knowledge graphs. The pipeline is designed to operate under conditions of label scarcity, multimodal heterogeneity, and real-world clinical uncertainty.

5.1 Overview of the Proposed Framework

Figure 5.1 summarises the full modelling pipeline. The framework proceeds through the following major stages:

1. Firstly, we preprocess the data and do a multimodal feature extraction from SNP, ECG, and biomarker domains.
2. Then, a tiered pseudo-label construction is conducted on the preprocessed data to handle initial sparse and uncertain cardiovascular diagnostic information accordingly.

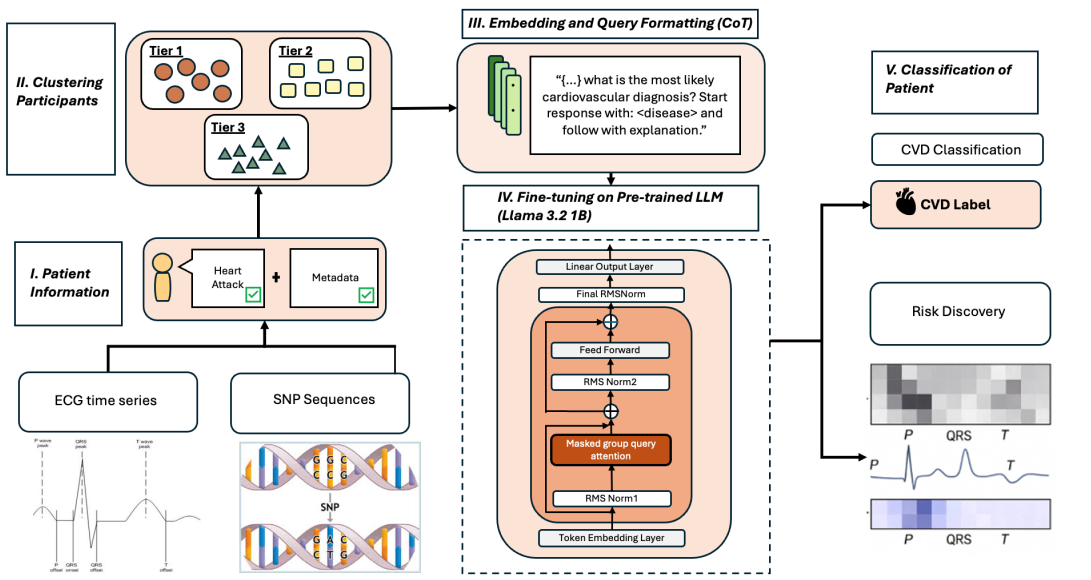

Figure 5.1: Overall Pipeline

3. Then, we use parameter-efficient LLMs, classical classifiers, chain of thought prompts and knowledge-graph augmentation for a multimodal model development.

4. The evaluation is predominantly conducted on label prediction performance, interpretability, and longitudinal forecasting.

This will be discussed more in detail in the subsequent sections.

5.2 Multimodal Feature Engineering

5.2.1 Participant-level features

As mentioned in Section 4.2, we first extract SNP genotypes from PLINK-formatted files and apply standard SNP quality control metrics such as MAF, HWE and GP to filter for rare variants while retaining vital SNPs.

Similarily, as mentioned in Section 4.3, ECG features were extracted from three annual

visitations per participant. Features included PR, QRS, QT, and QTc intervals, RR interval variability and derived pacing and repolarisation metrics.

Section 4.5 also mentions biochemical markers (HDL, LDL, triglycerides, glucose) and macronutrient intake, which were included to capture metabolic CVD risk. These markers serve as interpretable physiological indicators that complement genetic predisposition and ECG functional traits, but were only used in the context of LLMs and knowledge graph experiments.

5.3 Three-Tier Pseudo-Label Construction

As mentioned in the previous chapter 4, the Pheno.AI dataset is considered a few-labels environment solely due to the sparsely labeled medical-conditions dataset which acts as the main source of confidence labels. We address this few-labels problem by introducing a semi-supervised learning approach in the form of a three-tier label system. These tiers were explicitly extracted from a BioBERT model, which uses `021:medical-conditions` dataset from Pheno.AI and performs a semantic similarity on the labels that each participant was explicitly linked to.

Tier 1 cohort consists of participants with clear CVD diagnoses such as myocardial infarction or atrial fibrillation. Tier 2 cohort consists of participants cardiac-adjacent conditions including hypertension, arrhythmic symptoms, and so forth and the Tier 3 cohort represents the majority of the cohort consist of participants with no explicit cardiac annotation. The keywords used for the extraction and placement of each participant is summarized in the Table 5.1.

This structure enables us to perform weak supervision by assigning informative but flexible diagnostic categories, allowing models to learn both explicit and subverted patterns.

Even after applying quality control metrics, however, the genomic dataset remains extremely high-dimensional, with roughly 9.8 million SNPs preserved after QC (Ta-

Table 5.1: Keyword sets used for semi-supervised tier assignment

Tier	**Clinical Meaning**	**Associated Keywords**
Tier 1	Explicit CVD diagnoses	`coronary atherosclerosis, acute myocardial infarction, cardiac arrhythmia, pericarditis, atrial fibrillation, heart failure, myocarditis, complete atrioventricular block, cardiac arrest, endocarditis`
Tier 2	Cardiac-adjacent conditions	`hypertension, mild valve disease, left ventricular hypertrophy, sinus bradycardia, right bundle branch block, irregular heartbeat`
Tier 3	No explicit cardiac annotation	—

ble 4.2). This volume far exceeds what is informative or computationally practical for downstream modelling, as only a small fraction of variants per participant contribute meaningfully to cardiovascular risk. To mitigate this, we implement a tier-dependent SNP selection procedure that strategically narrows the variant space using a combination of biological relevance, GWAS evidence, and tier-specific phenotype associations.

5.3.1 Tier 1: GWAS Catalog–Driven SNP Extraction

Tier 1 contains participants with confirmed cardiovascular disease (CVD), making it the group with the strongest phenotype certainty. To identify relevant genetic variants per participant, we first extracted condition-specific labels filtered to Tier 1 diagnoses. For each label, we retrieved corresponding curated genome-wide association studies (GWAS) and expert-reviewed SNP lists, focusing on variants with established associations to the given condition. Specifically, the catalogs were taken from the ebi GWAS repository [**?**] and electing SNPs that showed genome-wide significant associations with a $p \leq 5 \times 10^{-8}$ with the condition. This manual curation step ensures that only biologically relevant and high-confidence SNPs (identified by rsIDs) are selected,

providing a consistent and interpretable genetic signature for each individual.

These condition-specific SNP sets serve two key purposes in our modeling pipeline. First, they ground the participant's genotypes in a clinically meaningful context, helping to reduce noise introduced by irrelevant variants. Second, they function as structured priors for downstream large language models (LLMs), which are not inherently pre-trained on genetic data. By injecting disease-associated variants into the prompt construction, particularly when designing chain-of-thought reasoning by guiding the LLM to focus on the genotype-condition based knowledge..

5.3.2 Tier 2: Hybrid SNP Selection Using Curated Variants + TF-IDF

Tier 2 corresponds to participants with cardiac-adjacent or metabolic conditions (e.g., hypertension). Because these phenotypes are less specific than Tier 1, we use a hybrid strategy that combines curated domain knowledge for known condition GWAS catalogs (for example, for hypertension), and with participant-level TF-IDF ranking.

Choppara et al. (2025) offers a method for TF-IDF extraction for relevant SNPs [CL25], where each SNP is first converted to a text token of the form `rsID:dosage`. The text document is determined as either 0, 1 or 2 (with 2 being the rarest). TF-IDF is then applied across all participants to identify discriminative variants (high IDF), and participant-specific variants (high TF). The top-k TF-IDF SNPs form the initial candidate set, thus allowing us to extract SNPs that are meaningful to each potential candidate with minimum bias.

5.3.3 Tier 3: Unsupervised SNP Clustering Using K-Means

Tier 3 comprises participants without any known cardiac diagnoses or labeled phenotypes. To extract a meaningful structure from this unlabeled cohort, we first construct TF-IDF representations over both SNP and ECG-derived features across all Tier 3

individuals. These embeddings are then subjected to unsupervised clustering, following approaches such as those outlined in [**?**], to uncover latent genotype-phenotype groupings between individual metrics.

To assess clinical relevance, each resulting cluster is also analyzed post hoc by comparing the distribution of Tier 1 SNPs and ECG biomarkers within the cluster. This enables us to infer putative risk levels. For example, participants who genetically resemble Tier 1 individuals yet currently exhibit only normative ECG signals may be annotated with a "future-risk" pseudo-label in the cluster. This strategy allows for the enrichment of the training dataset with soft labels that reflect potential subclinical cardiovascular risk.

Following the integration of all three tiers, we subsequently generated chain-of-thought prompts per participant, incorporating the relations between genotypic signals to cardiac diseases to improve model interpretability and risk stratification.[1].

5.4 Model Architecture

5.4.1 Chain-of-Thought Reasoning

A problem that we had to tackle was the idea of any generic models not having domain specific knowledge (which has been solved by the SNP/ECG extraction using tiers obviously), and their explicit lack of connection to such information. For this reason, we considered using a chain of thought reasoning approach which explicitly draws such a connection in a semi-supervised manner, thus allowing these models to learn from the connection. Each model was instructed to explain how SNPs influence cardiac electrophysiology, interpret ECG abnormalities, link biomarkers to metabolic CVD risk, and justify the final tier prediction.

This enabled generation of clinically meaningful rationales alongside predictions, improving interpretability and biological grounding. The CoT prompting can be summa-

[1]This clustering is later enhanced by knowledge graphs, described more in Section 5.5

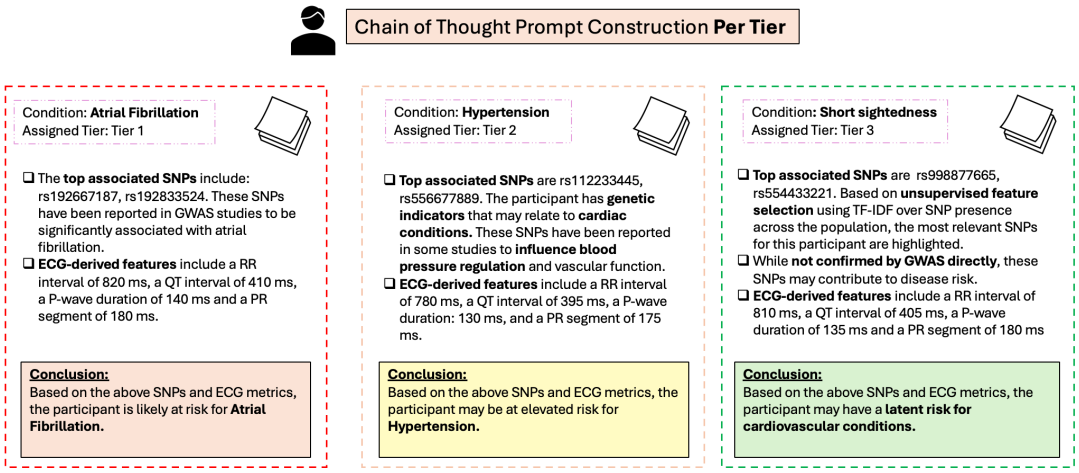

Figure 5.2: CoT Prompt Construction Across Different Tiers

rized in the figure 5.2 below.

By using this chain of thought prompt, thus the models can learn from ECG phenotypes and genomic information. However, as previously mentioned, it was not enough to just look at the two of them and use only k-means clustering. So, we enhance the participant profile using knowledge graphs, that are enhanced with biomarker information.

5.5 Knowledge Graph Integration

A lightweight clinical knowledge graph was constructed using biomarker relationships, genomic loci, and ECG traits. The KG was used to enrich LLM prompts with biomedical priors, identify clusters of cardiometabolic abnormalities (e.g., low-HDL and high-TG) and serve as a guide especially for Tier 3 inference, where labels are absent.

The edges were defined based on known lipid pathways, electrophysiological mechanisms, and gene-trait associations, and the general pattern is discussed more in Section 6.4.

Using this approach, we enhanced the chain of thought prompts, as shown in Figure 5.3. The difference here is that we incorporate biomarker based information and explicitly

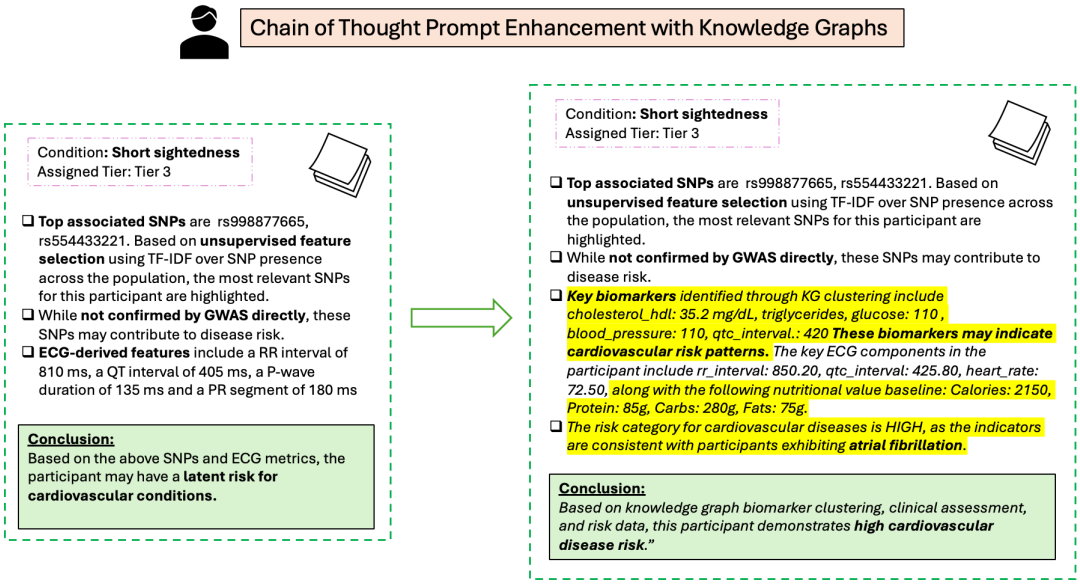

Figure 5.3: Knowledge Graph Enhanced Chain of Thought Prompt for a Tier 3 Participant.

state that these biomarkers were also found in a participant that had already been diagnosed with atrial fibrillation.

The knowledge graph construction surfaced with several biologically consistent risk patterns from the integrated SNP, ECG, and biochemical features. These included low HDL cholesterol, which is strongly associated with increased cardiovascular risk, as supported by ESC clinical reviews [GE20]. Elevated triglyceride levels, which is a well-established contributor to atherosclerosis and long-term cardiac complications [XYX+25]. High glucose levels, reflecting impaired metabolic regulation, which is tightly linked to cardiovascular disease progression [PLP+22]. Prolonged QTc interval, indicating delayed ventricular repolarisation, a known ECG marker associated with arrhythmias and elevated cardiac risk [noa22]. These relationships emerging from the KG confirm that the system is capturing realistic and clinically relevant connections between genetic, metabolic, and electrophysiological factors in cardiovascular disease. Figure 5.4 visualizes these trends accordingly.

Figures 5.5 and 5.6 depict tier trajectories over three years. Most participants remained consistently unlabeled (144), while several transitioned between tiers, primarily toward

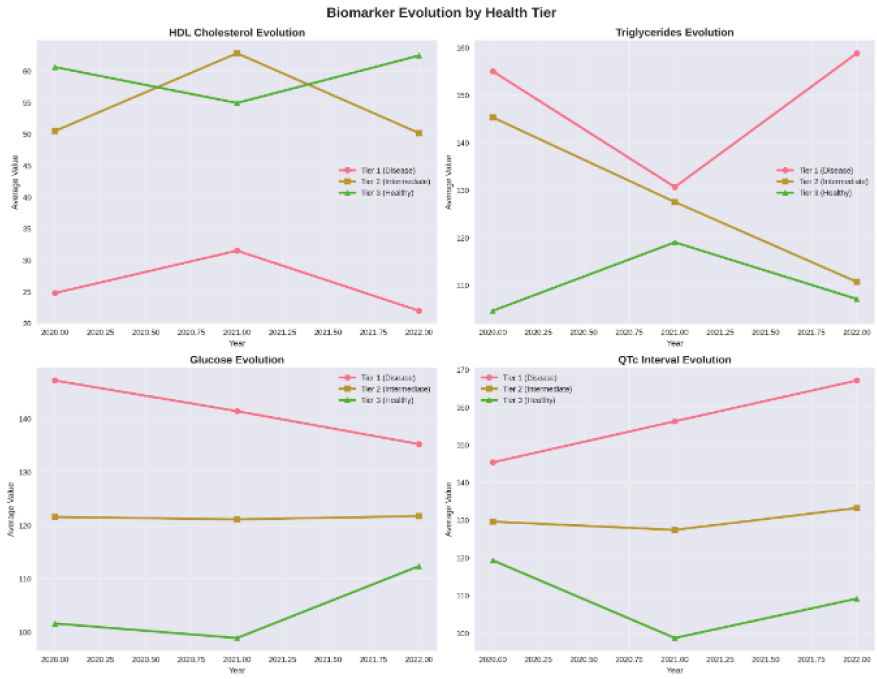

Figure 5.4: KG-identified biomarkers (apart from SNPs) that may contribute to cardiac risk.

Tier 2, reflecting cardiac-like symptoms. Only one participant moved from Tier 3 to Tier 1 and maintained this status for two years. These patterns highlight the importance of longitudinal modeling to capture early indicators of cardiovascular risk that static analyses might miss.

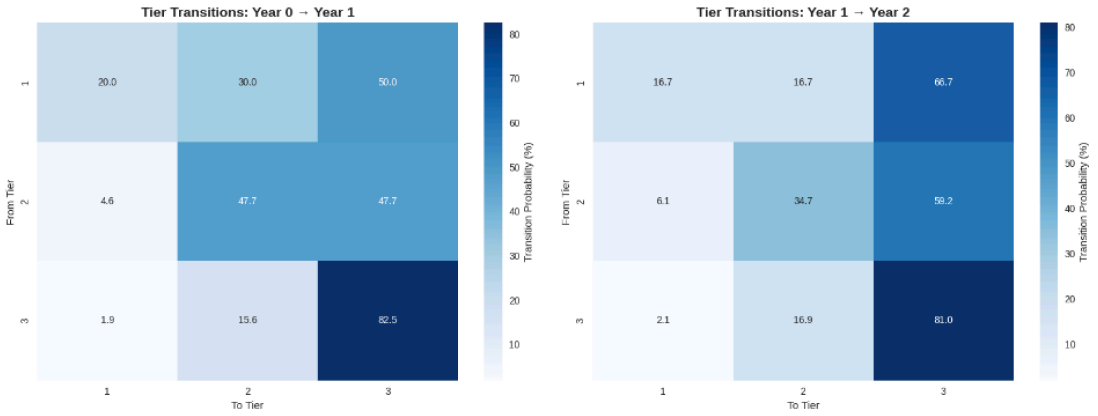

Figure 5.5: The tier transitions for each participant containing the visitation across three years (250 participants).

Following the construction of chain of thought prompt, we used a baseline traditional machine learning model approach and contrasted this with large language models to observe the classification power for each of these approaches. We also train large language models separately with knowledge-graph enhanced chain of thought prompts to showcase the true power of rich clinical context per participant cohort.

5.5.1 Machine Learning baseline and Large Language Models

The traditional machine learning classifiers that were were trained on concatenated on SNP and ECG features include: Logistic regression, Random forests, SVM, KNN,

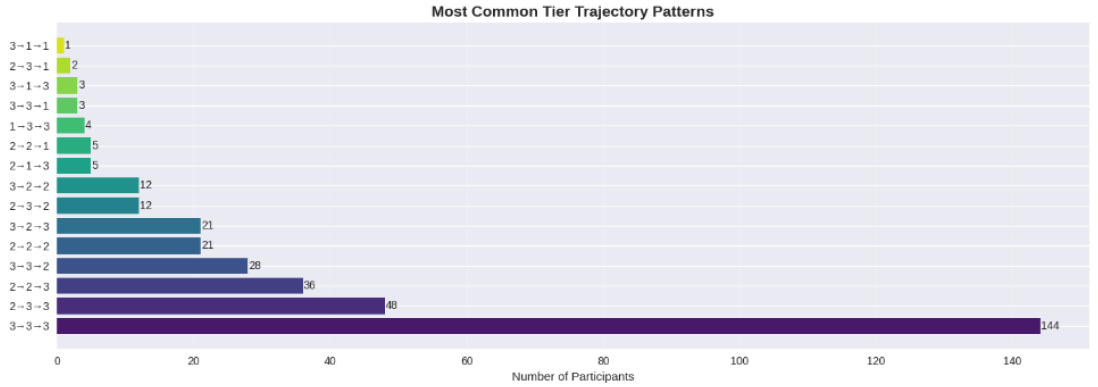

Figure 5.6: Most common tier trajectory patterns across three years, where most participants tend to not see any effect of CVDs across years.

HistGradientBoosting and LightGBM. These models provide interpretable baselines and highlight the added value of LLM-based reasoning.

Three parameter-efficient LLMs were considered for fine-tuning the chain of thought prompts: namely GPT-2 (124M parameters), Llama 3.2 1B and DeepSeek 1.3B.

5.6 Training Procedure

For each model, whether LLM-based or classical, the dataset was evaluated using 5-fold cross-validation. In the case of LLMs, greedy decoding was employed to generate deterministic outputs, which were subsequently post-processed to extract predicted tier labels. Chain-of-Thought (CoT) explanations were assessed qualitatively for clinical consistency; however, only the predicted labels were quantitatively matched using semantic similarity. In addition, experiments were conducted to forecast participant tiers in the following year, based on prior-year clusterings and data from participants with multiple visitations.

Model performance was assessed using the traditional classification report metrics such as accuracy and f1 scores, and was tested across tiers. These metrics capture both predictive performance and interpretability, essential for biomedical AI applications.

5.7 Summary

In summary, we propose to integrate multimodal feature engineering, semi-supervised tier construction, LLM-based reasoning, and clinical knowledge graphs into a unified pipeline for CVD risk stratification. This framework is specifically designed to operate in few-label settings, leveraging structured biological priors and Chain-of-Thought explanations to enhance interpretability and diagnostic reliability.

Chapter 6

Results and Discussion

This chapter presents the experimental results and observations from the proposed multimodal pipeline, which integrates SNP data, ECG-derived phenotypes, clinical biomarkers, and tier-based cardiac labels. Results are organized into the following components: firstly, the ablation experiments are used to justify the importance of describing the link between SNPs and ECGs, followed by traditional machine learning baselines, large language models (LLMs), and knowledge-graph-enhanced reasoning and concludes with temporal risk prediction analyses. Overall, the results highlight the strengths and limitations of each approach in addressing the few-label classification challenge in the Pheno.AI dataset.

6.1 Ablation Study: Feature Contributions

Prior to observing machine learning models, an ablation study was conducted to compared three configurations: (i) the baseline with combined SNP–ECG features and pseudo-labels, (ii) without ECG inputs, and (iii) without SNP inputs. This isolates each modality's contribution and highlights the value of multimodal fusion for cardiovascular risk prediction.

As shown in Table 6.1, removing either modality leads to a significant performance drop

across all metrics, confirming that the fusion of genomic and ECG features contributes synergistically to cardiovascular risk prediction. Notably, the models that do not consist of ECG inputs exhibited greater declines in recall, suggesting that dynamic cardiac signals play a key role in identifying subtle risk indicators that static SNP embeddings alone cannot capture. Conversely, removing SNPs led to reduced precision, indicating that genotype information enhances the model's ability to differentiate true positives from confounders. These findings demonstrate the value of multimodal learning for few-label biomedical prediction tasks, and allow us to justify the reason why we have to draw an explicit connection between the SNPs and ECGs.

Table 6.1: Ablation study comparing model performance under a multimodal (SNP + ECG) versus a unimodal (genotype only or phenotype only) configuration. The best results for each metric and the overall best model across all LLMs are in bold.

Model	**Type of Split**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
GPT-2	Baseline	0.810	0.822	0.840	0.830
	No ECG	**0.710**	**0.703**	**0.702**	**0.700**
	No SNP	0.740	0.754	0.740	0.745
DeepSeek 1.3B	Baseline	0.920	0.831	0.810	0.820
	No ECG	**0.810**	**0.791**	**0.724**	**0.746**
	No SNP	0.822	0.711	0.723	0.714
Llama 3.2 1B	Baseline	0.920	0.830	0.891	0.840
	No ECG	**0.702**	**0.691**	**0.650**	**0.644**
	No SNP	0.755	0.722	0.724	0.726

6.2 Traditional Machine Learning Results

To establish baseline performance prior to evaluating large language models, six traditional machine learning classifiers were trained on TF-IDF SNP representations and ECG-derived phenotypes. The models included Logistic Regression, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbour (KNN), HistGradientBoosting, and LightGBM. We wanted to use a few labels setup to explore the machine learning models, so we did not imply any synthetic oversampling techniques such as SMOTE. To mimic a realistic few-label environment and ensure that the traditional ML models did not over-generalize across many labels, we only focused on three primary cardiac

conditions: atrial fibrillation, coronary atherosclerosis and hypertension, with all the other participants labeled as "Other".

Table 6.2 provides a high-level summary and comparison of accuracy, macro-F1, and weighted-F1 across all models for the purposes of our explanation. As shown, although overall accuracy remains deceptively high across all methods, the macro-F1 scores reveal a consistent inability to correctly classify minority cardiovascular conditions. The full classification reports for each model are provided in Appendix A.

Table 6.2: Summary of Traditional Machine Learning Performance Across Models

Model	**Accuracy**	**Macro F1**	**Weighted F1**
Logistic Regression	0.92	0.25	0.90
Random Forest	0.96	0.67	0.90
SVM	0.93	0.33	0.93
KNN	0.93	0.63	0.93
HistGradientBoosting	0.97	0.39	0.91
LightGBM	0.93	0.25	0.90

6.2.1 Model-Specific Observations

Logistic Regression. Achieved high accuracy (0.92) but produced zero recall for atrial fibrillation and coronary atherosclerosis. This demonstrates an inability for Logistic Regression to learn meaningful boundaries in sparse SNP space.

Random Forest. Performed the best among baseline models, particularly for atrial fibrillation (F1 = 0.70). However, it still exhibited major drops in recall for other minority classes due to overfitting to the majority distribution.

Support Vector Machine (SVM). SVM achieved high precision but extremely low recall for minority classes, indicating an overly conservative decision boundary in high-dimensional feature space.

K-Nearest Neighbour (KNN). KNN showed more balanced performance across classes but suffered from the curse of dimensionality, causing unstable predictions for hypertension and coronary conditions.

HistGradientBoosting and LightGBM. While achieving the highest accuracy (0.97 and 0.93), these models completely failed to detect minority cardiac classes, highlighting their strong majority-class bias.

In our observations, we found a consistent pattern that each of the classifiers follow while identifying the labels based on the SNP and ECG information. While the overall accuracy remained high (92-97%), the performance on minority cardiovascular classes was extremely poor. This reflects structural challenges inherent to the task as the severe class imbalance caused most models to collapse toward the majority "Other" class. Similarily, sparse, high-dimensional SNP TF-IDF vectors negatively impacted linear and distance-based models due to no context and non-linear genotype-phenotype relationships were poorly captured by traditional models.

These limitations resulted in very low macro-F1 scores even for tree-based models, despite deceptively high accuracy.

Therefore, traditional ML methods provide a reliable baseline but fail to capture the complex, multimodal relationships required for robust cardiovascular risk prediction. Their inability to model sparse genomic signals, combined with extreme class imbalance, motivates our preference and transition to large language models, which can incorporate biological priors, context, and non-linear interactions.

6.3 Large Language Model (LLM) Results

As descibed in Chapter 5, three Large Language Models (namely, DeepSeek 1.3B [GZY+24], LLaMA 3.2 1B [Met], and GPT-2) were fine-tuned using Low-Rank Adaptation (LoRA) [HSW+21] and Chain-of-Thought (CoT) prompting. These models were evaluated both under a

Few-Labels setting (a set of 350 participants randomly selected per tier) and a full Skyline setting (all 10160 participants). Across both regimes, LLMs demonstrated substantial improvements over traditional machine learning baselines, particularly in recall and F1-score for cardiac-relevant tiers.

Table 6.3 provides an overview of overall performance across all participants.

Table 6.3: Overall performance comparison of LLMs under a few labels setting and skyline overview. The best results for each metric and the overall best model across all LLMs are in bold.

Model	**Accuracy**	**Precision**	**Recall**	**F1**
		Few Labels [1050 participants]		
GPT-2	0.811	**0.890**	0.812	0.842
LLaMA-3.2 1B	0.880	0.822	0.790	0.790
DeepSeek 1.3B	**0.892**	0.860	0.840	0.832
		Skyline [Full Data]		
GPT-2	0.800	0.810	0.809	0.810
LLaMA-3.2 1B	0.901	0.832	0.780	0.790
DeepSeek 1.3B	0.910	0.869	0.830	0.840

6.3.1 Overall Performance Trends

Across all metrics, LLMs consistently outperformed traditional models. DeepSeek 1.3B achieved the highest overall accuracy (0.892 in Few-Labels, 0.910 in Skyline), while GPT-2 showed competitive precision under Few-Labels constraints. LLaMA 3.2 1B performed robustly across both data regimes with stable precision–recall trade-offs.

While traditional ML models often collapsed to the majority "Other" class, LLMs captured richer non-linear interactions between SNP representations, ECG-derived features, and participant metadata. CoT prompting further enhanced interpretability, enabling models to explicitly reference QTc prolongation, PR interval deviation, or lipid biomarkers when generating cardiac risk predictions.

6.3.2 Tier-Wise Performance

To better understand the behaviour of LLMs across different levels of phenotype certainty, we would like to present the model evaluations for Tiers 1–3. Tier definitions follow Section 5.3, with Tier 1 capturing strongly labelled cardiac cases, Tier 2 representing cardiac-adjacent individuals, and Tier 3 corresponding to largely unlabelled or weakly labelled participants.

Tables 6.4, 6.5 and 6.6 summarizes the model performance across these partitions under both evaluation regimes, and will be discussed in the following section.

Table 6.4: Performance for Tier 1 participants under a few labels setting and skyline overview. The best results for each metric and the overall best model across all LLMs are in bold.

Model	**Accuracy**	**Precision**	**Recall**	**F1**
		Few Labels [350 participants]		
GPT-2	0.810	0.822	0.840	0.830
LLaMA-3.2 1B	**0.920**	0.830	**0.891**	**0.840**
DeepSeek 1.3B	**0.920**	**0.831**	0.810	0.820
		Skyline [Full Data]		
GPT-2	0.820	0.825	0.845	0.835
LLaMA-3.2 1B	0.925	0.840	0.895	0.855
DeepSeek 1.3B	0.935	0.870	0.835	0.845

Table 6.5: Performance comparison for Tier 2 participants under a few labels setting and skyline overview. The best results for each metric and the overall best model across all LLMs are in bold.

Model	**Accuracy**	**Precision**	**Recall**	**F1**
		Few Labels [350 participants]		
GPT-2	0.800	0.813	0.791	0.800
LLaMA-3.2 1B	0.890	0.824	**0.820**	0.822
DeepSeek 1.3B	**0.910**	**0.850**	**0.820**	**0.830**
		Skyline [Full Data]		
GPT-2	0.810	0.820	0.800	0.810
LLaMA-3.2 1B	0.905	0.835	0.825	0.830
DeepSeek 1.3B	0.925	0.860	0.830	0.840

Table 6.6: Performance comparison for Tier 3 participants under a few labels setting and skyline overview. The best results for each metric and the overall best model across all LLMs are in bold.

Model	Accuracy	Precision	Recall	F1
		Few Labels [350 participants]		
GPT-2	0.811	**0.890**	0.812	**0.842**
LLaMA-3.2 1B	0.880	0.822	0.790	0.790
DeepSeek 1.3B	**0.892**	0.860	**0.840**	0.832
		Skyline [Full Data]		
GPT-2	0.820	0.895	0.815	0.850
LLaMA-3.2 1B	0.890	0.830	0.795	0.805
DeepSeek 1.3B	0.900	0.870	0.845	0.855

6.3.3 Tier-Specific Insights

Tier 1 (High-Confidence Labels). All three LLMs demonstrated strong performance in this Tier, leveraging clear diagnostic labels and well-defined ECG patterns. DeepSeek 1.3B and LLaMA 3.2 1B achieved accuracies above 0.92 in both Few-Labels and Skyline evaluations. CoT analyses show that models frequently referenced QTc prolongation, RR irregularity, and SNP-derived risk signatures associated with arrhythmogenic pathways.

Tier 2 (Cardiac-Adjacent). Performance decreased slightly due to label noisiness and weaker signals. However, DeepSeek 1.3B remained robust (Few-Labels F1 = 0.830), effectively integrating semi-relevant SNP clusters extracted via TF–IDF in combination with ECG variation. LLaMA 3.2 1B also demonstrated consistent precision and recall across both settings.

Tier 3 (Unlabelled Participants). Despite the absence of strong labels, LLMs maintained coherent predictions by relying on ECG morphology and participant-dependent cardiac trends. DeepSeek and GPT-2 performed particularly well, suggesting that CoT reasoning aids extrapolation under limited supervision.

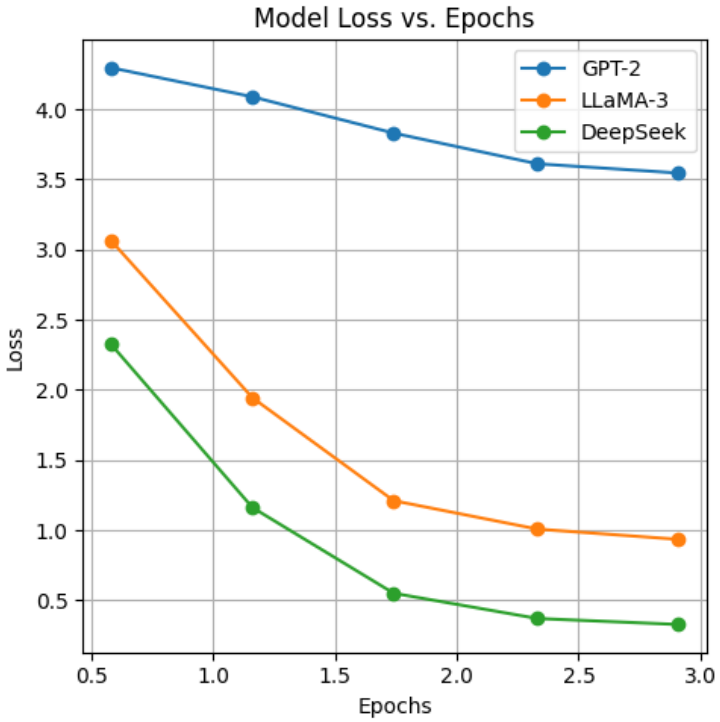

Figure 6.1: Training Loss Plotted for All Three LLMs.

6.3.4 Training Behaviour

Training loss curves (Figure 6.1) further support the results by showcasing a smooth and stable convergence for all models. LoRA regularisation reduced overfitting, while early stopping ensured robustness in both the Few-Labels and full Skyline training conditions. We can also observe the general trend that each LLMs follow; DeepSeek converges the quickest, followed closely by Llama 3.2 1B. However, GPT-2 lags behind due to being a smaller and slower model.

While these results demonstrate that LLMs substantially outperform traditional machine learning models and provide stronger predictive power across all tiers, an important limitation still remains. As established in Section 5.3.3, K-Means clustering is useful as an initial exploratory tool, but it only offers a coarse grouping of participants. It does not incorporate the underlying biological structure present in SNP networks, biomarker interactions, or cardiac physiology.

To meaningfully enhance participant profiling, especially for Tier 3, where labels are sparse, there is a need for a more detailed representation of multimodal biological relationships. This motivates the use of *knowledge graphs* (KGs), which allow us to explicitly model relationships between genetic variants, biomarkers, ECG-derived features, and clinical risks. The next section presents the results of applying KG-based clustering and analysis to the dataset.

6.4 Knowledge Graph (KG) Results

Knowledge graphs integrate biological context by representing SNP–gene–trait interactions, biomarker dependencies, and ECG-derived phenotypes as structured relational networks. Using the KG-based clustering pipeline introduced in Chapter 5, we examined whether the inclusion of biological priors improves risk stratification, particularly in the unlabeled Tier 3 population.

KG clustering produced clearer separation of metabolic and cardiac subgroups compared to K-Means alone and revealed interpretable associations, such as reduced HDL concentrations clustering with elevated cardiovascular risk, prolonged QTc intervals appearing in subclusters linked to arrhythmogenic SNP pathways, triglyceride-rich metabolic phenotypes grouping with hypertension-linked ECG irregularities.

The LLMs are separately trained on the knowledge-graph enhanced Chain of Thought Prompts and evaluated with the accuracy, precision, recall and f1 metrics, and trained on the same cohort in the few-labels environment as the baseline LLM models presented in Section 6.3.3. Table 6.7 represents these results. Compared to Table 6.6 for a baseline non-KG enhanced clustering, it can be seen that the precision and recall are significantly increased due to more context and weight to the cluster meaning, thus confirming that using biomarkers for better context can significantly improve model context.

Table 6.7: Tier 3 Performance with Knowledge Graph (KG) Augmentation

Model	Accuracy	Precision	Recall	F1 Score
Without KG Augmentation				
GPT-2	0.811	0.890	0.812	0.842
DeepSeek 1.3B	0.892	0.860	0.840	0.832
Llama 3.2 1B	0.880	0.822	0.790	0.790
With KG Augmentation				
GPT-2	**0.851**	**0.900**	0.820	0.842
DeepSeek 1.3B	**0.900**	0.890	**0.840**	**0.832**
Llama 3.2 1B	**0.890**	**0.940**	0.800	**0.850**

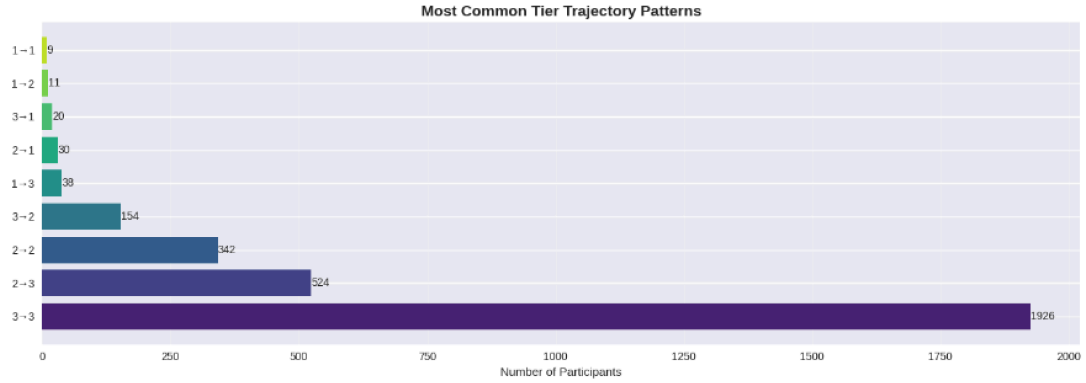

Figure 6.2: Most common tier trajectory patterns across two years, where most participants tend to not see any effect of CVDs across years.

6.5 Temporal Risk Prediction

Following the classification, we also use KG-enhanced clustering for forecasting.

Figure 6.2 showcases the participant split across two years. Very similar to the KG-based dataset information given in Section **??**, we can infer that majority of participants tend to not contract any cardiovascular diseases over two years time, with some of them even moving from latent cardiovascular risks based tiers (Tier 2) to Tier 3. However, we can also see that 154 participants move from Tier 3 to 2, while 20 participants move from Tier 2 to 1, making it a significant change worth investigating.

For this reason, we used DeepSeek 1.3B to forecast Year 2 risk tiers based solely on Year 0 baseline data. The analysis focused on a stratified subset comprising 20% of

Table 6.8: Risk Prediction Using Knowledge Graphs (using DeepSeek 1.3B)

	Precision	**Recall**	**F1-Score**	**Support**
Tier 1	0.60	0.40	0.48	59
Tier 2	0.90	0.87	0.88	255
Tier 3	0.93	0.95	0.94	297
Accuracy			0.92	611
Macro Avg	0.81	0.74	0.77	611
Weighted Avg	0.92	0.92	0.92	611

the original dataset, restricted to participants with only two visitation records. Risk tier assignments were determined via keyword extraction, where predictions containing terms such as "high", "Tier 1", or "confirmed" were classified as Tier 1, while analogous terms mapped to Tiers 2 and 3. The results, summarized in Table 6.8, reveal distinct performance patterns across tiers:

Tier 3 (No Explicit Cardiac Relevance) The model achieved exceptionally high precision (0.93) and recall (0.95), with an F1 score of 0.94. This strong performance reflects the stability and clarity of patterns in Tier 3, where the majority of participants exhibit consistent ECG-genomic relationships. The model reliably captured dominant features in this group, demonstrating robust generalization for low-risk profiles.

Tier 2 (Cardiac-Adjacent Conditions) Performance remained solid, with precision and recall in the high 80s (0.90 and 0.87, respectively). This is particularly encouraging, as Tier 2 comprises participants with indirect or ambiguous cardiac associations. The inclusion of temporal context appeared to stabilize the model's interpretations, allowing it to detect subtle progression signals over time.

Tier 1 (High-Confidence Cardiac Diagnoses) Predicting Tier 1 cases proved most challenging, as expected. These cases are rare and often lack consistent temporal patterns, leading to moderate precision (0.60) and lower recall (0.40). However, the model still identified meaningful Tier 1 cases, particularly those supported by

GWAS-validated SNPs. The relative difficulty in this tier aligns with real-world clinical datasets, where explicit diagnoses are sparse and heterogeneous.

While Tier 1 prediction remains difficult to predict over time, which is a reflection of its scarcity and variability, both Tier 2 and Tier 3 exhibited clear performance gains. The overall accuracy of 92% underscores the model's ability to generalize across the majority of cases, with particularly strong results for stable (Tier 3) and moderately ambiguous (Tier 2) subgroups. The findings highlight the potential of temporal risk stratification in few-label settings.

6.6 Summary of Results

Thus, our research highlights key insights into the performance of traditional machine learning models, large language models (LLMs), and knowledge-graph-enhanced reasoning for cardiovascular disease classification in few-label settings. By integrating baseline SNPs, ECG trends, and biomarker trajectories, the model effectively mimics clinical reasoning, where progression over time is a critical diagnostic indicator. The Traditional models—Logistic Regression, Random Forest, and SVM—struggled with high-dimensional, sparse SNP data and class imbalance: while overall accuracy was high (92–97%), macro-F1 scores remained below 0.70, reflecting poor classification of minority cardiac conditions. LLMs (DeepSeek 1.3B, LLaMA 3.2 1B, GPT-2) outperformed these baselines by leveraging Chain-of-Thought prompting and multimodal inputs (SNPs, ECG, biomarkers), achieving up to 92%accuracy in few-label scenarios and 93.5% with full data, while providing interpretable predictions that link genetic and physiological cues. Knowledge-graph (KG) augmentation further improved results, particularly for Tier 3 participants, encoding biological relationships among SNPs, biomarkers, and ECG features to yield a 5% absolute F1-score gain and revealing clinically consistent patterns, such as low HDL correlating with cardiovascular risk. Temporal risk prediction experiments also demonstrated the feasibility of forecasting Year 2 cardiac risk using Year 0 data, achieving 93% precision and 95% recall for Tier 3, though Tier 1 remained challenging due to data sparsity. These results emphasize

the potential of multimodal LLMs and KGs for complex, few-label classification tasks, while highlighting the need for further refinement in rare, high-risk cases.

6.7 Limitations

Despite the promising results, however, we faced certain limitations, as discussed below:

- Resource Constraints: There were workspace and resource limitations, such as limited GPU memory on A10G GPUs, which had required us to scale down participant cohorts for manageable runs. This may have impacted the ultimate performance of the models, particularly in comparison to state-of-the-art biomedical LLMs trained on larger datasets.
- Model Size: The model sizes used in this study were relatively small compared to cutting-edge biomedical LLMs. Larger models with more parameters might achieve even better performance but require significantly more computational resources.
- API Restrictions: There were workspace restrictions on using hosted models, such as those from HuggingFace or GPT-based models, limited the study to locally downloaded models and checkpoints for training. This constraint may have affected the diversity and robustness of the models evaluated.
- Clinical Validation: While the results are promising, clinical validation and real-world testing are essential before these models can be deployed in healthcare settings. Collaborations with medical professionals and further validation studies are needed to ensure the reliability and safety of the proposed approaches.

These limitations can be addressed in future research, which could involve scaling up the participant cohorts, exploring larger and more diverse models, and conducting extensive clinical validation studies. Additionally, integrating multimodal data sources

and advancing temporal risk prediction pipelines could further enhance the utility of these models in cardiogenomic diagnosis.

Chapter 7

Conclusion and Future Work

In conclusion, through this thesis we introduce a multimodal, few-label framework for cardiovascular disease (CVD) risk stratification, integrating SNP genotypes, ECG-derived features, and structured diagnostic labels. The study integrates traditional machine learning, parameter-efficient LLMs, and knowledge-graph reasoning to capture complex genotype–phenotype interactions while having limited cardiac annotations. Results show that interpretable multimodal LLMs can uncover clinically meaningful signals even in data-constrained settings.

Our primary contribution is a few-label multimodal LLM framework capable of jointly modelling genomic and physiological modalities for cardiac risk prediction. By constructing a three-tier pseudo-labeling system, models not only learn from explicit diagnoses (Tier 1) but also from cardiac-adjacent or unlabeled cohorts (Tiers 2-3), facilitating discovery of latent risk profiles. The use of Chain-of-Thought (CoT) prompting provided structured, clinically grounded explanations linking predictions to ECG irregularities, biomarker abnormalities, and SNP associations.

Among the pre-trained large language models of GPT-2, Llama 3.2 1B, and DeepSeek 1.3B, DeepSeek achieved the highest performance. Ablation studies confirmed that removing either SNP or ECG modalities degraded results, highlighting the importance of multimodal fusion. Knowledge-graph reasoning further improved predictions, particularly

for Tier 3 participants, by incorporating biomarker clusters (e.g., low HDL, elevated triglycerides, high glucose, prolonged QTc) as structured priors. Longitudinal analyses demonstrated meaningful tier transitions, illustrating the framework's potential for temporal risk prediction.

Future work that will be conducted under improved resource conditions could address this constraint appropriately. While we deliberately employed smaller language models to assess feasibility, scaling to larger architectures such as Llama-3 7B[AI@24] or GPT-4[Ope24] could substantially enhance performance, particularly in Tier 3 scenarios or zero-shot classification tasks. Causal language models lack native classification heads, making outputs sensitive to prompt structure; semantic similarity metrics were employed to align predictions with known cardiac conditions. Moving forward, however, post-processing strategies will be incorporated to constrain output variability and enforce label consistency, particularly important in clinical forecasting tasks, where the interpretation of subclinical risk must remain robust and standardized across decoders like Llama.

A version of this research has been accepted at the IEEE BIBM LLMBDA conference and submitted to AAAI 2026, with the pre-print versions available on arXiv: `https://arxiv.org/abs/2508.07127` [MFL$^+$25] and `https://arxiv.org/pdf/2510.16536` [MLF$^+$25].

Overall, this thesis demonstrates that few-label, multimodal, interpretable LLM frameworks provide a strong foundation for precision cardiovascular risk prediction, offering both robust performance and clinically meaningful transparency. These findings highlight the potential of explainable AI systems to support early detection and personalized cardiogenomic risk assessment in real-world limited labels biomedical settings.

Bibliography

[ABT20] Nikita Abramovs, Andrew Brass, and May Tassabehji. Hardy-Weinberg Equilibrium in the Large Scale Genomic Sequencing Era. *Frontiers in Genetics*, 11:210, March 2020.

[AI@24] AI@Meta. Llama 3 model card, 2024.

[BMBP19] Kwembe B.a, Aliyu Mohammed, J. G. Bashayi, and A. A. Patrick. Prediction of ECG Signals Using Feedforward Neural Networks. *International Journal of Current Innovations in Advanced Research*, pages 1–8, 2019.

[CL25] Prashanth Choppara and Bommareddy Lokesh. Leveraging quantum lstm for high-accuracy prediction of viral mutations. *IEEE Access*, 13:25282–25300, 2025.

[CMY+17] Ingrid Christophersen, Jared Magnani, Xiaoyan Yin, John Barnard, Lu-Chen Weng, Dan Arking, Maartje Niemeijer, Steven Lubitz, Christy Avery, Qing Duan, Joshua Bis, Kathleen Kerr, Aaron Isaacs, Martina Müller-Nurasyid, Kari North, Alex Reiner, Lesley Tinker, and Patrick Ellinor. Fifteen genetic loci associated with the electrocardiographic p waveclinical perspective. *Circulation: Cardiovascular Genetics*, 10:e001667, 08 2017.

[EBI] What are genome wide association studies (GWAS)? | GWAS Catalog.

[ESGA25] Santiago Espinosa-Salas and Mauricio Gonzalez-Arias. Nutrition: Macronutrient Intake, Imbalances, and Interventions. In *StatPearls*. StatPearls Publishing, Treasure Island (FL), 2025.

[GE20] 'Sadi Güleç' and 'Cetin Erol'. High-density lipoprotein cholesterol and risk of cardiovascular disease, 2020. Issue: 3 Volume: 19.

[GLG+] Gary H. Gibbons, Choong Chin Liew, Mark O. Goodarzi, Jerome I. Rotter, Willa A. Hsueh, Helmy M. Siragy, Richard Pratt, and Victor J. Dzau. Genetic markers. 109(25):IV–47. Num Pages: IV-58 Publisher: American Heart Association.

[GSL+25] Bowen Gu, Vivian Shao, Ziqian Liao, Valentina Carducci, Santiago Romero Brufau, Jie Yang, and Rishi J. Desai. Scalable information extraction from free text electronic health records using large language models. *BMC Medical Research Methodology*, 25(1):23, January 2025.

[GWA08] GWAS Catalog, 2008.

[GZY+24] Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Y. Wu, Y. K. Li, Fuli Luo, Yingfei Xiong, and Wenfeng Liang. DeepSeek-Coder: When the Large Language Model Meets Programming – The Rise of Code Intelligence. In *DeepSeek-Coder*. arXiv, January 2024. arXiv:2401.14196 [cs].

[HSW+21] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[KMIA24] Ahmed Akib Jawad Karim, Muhammad Zawad Mahmud, Samiha Islam, and Aznur Azam. Enhancing Multi-Class Disease Classification: Neoplasms, Cardiovascular, Nervous System, and Digestive Disorders Using Advanced LLMs, November 2024. arXiv:2411.12712 [cs].

[KS17] Santosh Kumar and Gadadhar Sahoo. A Random Forest Classifier based on Genetic Algorithm for Cardiovascular Diseases Diagnosis. *International Journal of Engineering, Transactions B: Applications*, 30:1723–1729, November 2017.

[LAH+18] Seung-Pyo Lee, Euan A. Ashley, Julian Homburger, Colleen Caleshu, Eric M. Green, Daniel Jacoby, Steven D. Colan, Edmundo Arteaga-Fernández, Sharlene M. Day, Francesca Girolami, Iacopo Olivotto, Michelle Michels, Carolyn Y. Ho, Marco V. Perez, and SHaRe Investigators. Incident Atrial Fibrillation Is Associated With MYH7 Sarcomeric Gene Variation in Hypertrophic Cardiomyopathy. *Circulation. Heart Failure*, 11(9):e005191, September 2018.

[Lee13] Dong-Hyun Lee. Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)*, 07 2013.

[LT19] Alberto Lorenzatti and Peter P. Toth. New Perspectives on Atherogenic Dyslipidaemia and Cardiovascular Disease. August 2019.

[maf] Minor Allele Frequency - an overview | ScienceDirect Topics.

[MBLA12] Ana Maggio, María Bonomini, Eric Laciar, and Pedro Arini. Quantification of Ventricular Repolarization Dispersion Using Digital Processing of the Surface ECG. January 2012.

[Met] Meta. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models.

[MFL+25] Niranjana Arun Menon, Iqra Farooq, Yulong Li, Sara Ahmed, Yutong Xie, Muhammad Awais, and Imran Razzak. How effectively can large language models connect snp variants and ecg phenotypes for cardiovascular risk prediction?, 2025.

[MKR+16] Bettina Mieth, Marius Kloft, Juan Antonio Rodríguez, Sören Sonnenburg, Robin Vobruba, Carlos Morcillo-Suárez, Xavier Farré, Urko M. Marigorta, Ernst Fehr, Thorsten Dickhaus, Gilles Blanchard, Daniel Schunk, Arcadi Navarro, and Klaus-Robert Müller. Combining Multiple Hypothesis Testing with Machine Learning Increases the Statistical Power of Genome-wide Association Studies. *Scientific Reports*, 6(1):36671, November 2016. Publisher: Nature Publishing Group.

[MLF+25] Niranjana Arun Menon, Yulong Li, Iqra Farooq, Sara Ahmed, Muhammad Awais, and Imran Razzak. Few-Label Multimodal Modeling of SNP Variants and ECG Phenotypes Using Large Language Models for Cardiovascular Risk Stratification, October 2025. arXiv:2510.16536 [q-bio].

[NCB98] Home - SNP - NCBI, 1998.

[NHS17] Cardiovascular disease, October 2017. Section: conditions.

[NHW+25] Vasudeva Reddy Netala, Tianyu Hou, Yanbo Wang, Zhijun Zhang, and Sireesh Kumar Teertam. Cardiovascular biomarkers: Tools for precision diagnosis and prognosis. *Int. J. Mol. Sci.*, 26(7):3218, March 2025.

[noa22] Arrhythmias - Long QT Syndrome | NHLBI, NIH, March 2022.

[Ope24] OpenAI. GPT-4 Technical Report. `https://arxiv.org/pdf/2303.08774`, 2024. [Accessed 04-08-2025].

[Phe21] participant_journey – Pheno.AI, 2021.

[PLP+22] Anastasia V Poznyak, Larisa Litvinova, Paolo Poggio, Vasily N Sukhorukov, and Alexander N Orekhov. Effect of glucose levels on cardiovascular risk. *Cells*, 11(19):3034, September 2022.

[PSZ22] Wenjuan Peng, Yuan Sun, and Ling Zhang. Construction of genetic classification model for coronary atherosclerosis heart disease using three machine learning methods. *BMC Cardiovascular Disorders*, 22(1):42, February 2022.

[RNG+23] Max Tigo Rietberg, Van Bach Nguyen, Jeroen Geerdink, Onno Vijlbrief, and Christin Seifert. Accurate and Reliable Classification of Unstructured Reports on Their Diagnostic Goal Using BERT Models. *Diagnostics*, 13(7):1251, March 2023.

[RPS22] Kandukuri Ratna Prakarsha and Gaurav Sharma. Time series signal forecasting using artificial neural networks: An application on ECG signal. *Biomedical Signal Processing and Control*, 76:103705, July 2022.

[SA24] Farnoush Shishehbori and Zainab Awan. Enhancing Cardiovascular Disease Risk Prediction with Machine Learning Models, February 2024. arXiv:2401.17328 [q-bio].

[SH24] Oluwafemi A. Sarumi and Dominik Heider. Large language models and their applications in bioinformatics. *Computational and Structural Biotechnology Journal*, 23:3498–3505, December 2024.

[WHO25] WHO. Cardiovascular diseases, 2025.

[WX25] You Wu and Lei Xie. AI-driven multi-omics integration for multi-scale predictive modeling of genotype-environment-phenotype relationships. *Computational and Structural Biotechnology Journal*, 27:265–277, January 2025.

[WY23] Junfan Chen Lihong Wang Jaein Kim Weiyi Yang, Richong Zhang. Prototype-Guided Pseudo Labeling for Semi-Supervised Text Classification — aclanthology.org. `https://aclanthology.org/2023.acl-long.904/`, 2023. [Accessed 25-11-2025].

[XGX+24] Tianhan Xu, Yixun Gu, Mantian Xue, Renjie Gu, Bin Li, and Xiang Gu. Knowledge graph construction for heart failure using large language models with prompt engineering. *Frontiers in Computational Neuroscience*, 18, July 2024. Publisher: Frontiers.

[XYX+25] Jian Xu, Chao Yu, Jiawei Xu, Vetle I Torvik, Jaewoo Kang, Mujeen Sung, Min Song, Yi Bu, and Ying Ding. PubMed knowledge graph 2.0: Connecting papers, patents, and clinical trials in biomedical science. *Sci. Data*, 12(1):1018, June 2025.

[Yar95] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196. Association for Computational Linguistics, 1995.

Appendix: Traditional Machine Learning Classification Reports

This appendix contains the full classification reports for all traditional machine learning models evaluated in this thesis. These tables complement the summary presented in Chapter **??**, providing detailed precision, recall, F1-scores, and class supports for each cardiovascular category.

A.1 Logistic Regression

Table A.1: Classification Report: Logistic Regression

	Precision	**Recall**	**F1-Score**	**Support**
Atrial Fibrillation	0.00	0.00	0.00	24
Coronary Atherosclerosis	0.00	0.00	0.00	90
Hypertension	0.09	0.03	0.08	2,494
Other	0.94	0.97	0.95	3,126
Accuracy			0.92	5,734
Macro Avg	0.25	0.25	0.25	5,734
Weighted Avg	0.88	0.92	0.90	5,734

A.2 Random Forest

Table A.2: Classification Report: Random Forest

	Precision	**Recall**	**F1-Score**	**Support**
Atrial Fibrillation	1.00	0.56	0.70	24
Coronary Atherosclerosis	1.00	0.21	0.40	90
Hypertension	0.92	0.38	0.54	2,494
Other	0.97	0.98	0.98	3,126
Accuracy			0.96	5,734
Macro Avg	0.97	0.54	0.67	5,734
Weighted Avg	0.96	0.96	0.90	5,734

A.3 Support Vector Machine

Table A.3: Classification Report: Support Vector Machine

	Precision	**Recall**	**F1-Score**	**Support**
Atrial Fibrillation	0.96	0.14	0.27	24
Coronary Atherosclerosis	1.00	0.04	0.08	90
Hypertension	0.97	0.03	0.10	2,494
Other	0.94	0.96	0.96	3,126
Accuracy			0.93	5,734
Macro Avg	0.99	0.30	0.33	5,734
Weighted Avg	0.94	0.94	0.93	5,734

A.4 K-Nearest Neighbour

Table A.4: Classification Report: K-Nearest Neighbour

	Precision	**Recall**	**F1-Score**	**Support**
Atrial Fibrillation	0.84	0.65	0.74	24
Coronary Atherosclerosis	0.23	0.35	0.24	90
Hypertension	0.69	0.49	0.55	2,494
Other	0.98	0.96	0.97	3,126
Accuracy			0.93	5,734
Macro Avg	0.68	0.62	0.63	5,734
Weighted Avg	0.95	0.95	0.93	5,734

A.5 HistGradientBoosting

Table A.5: Classification Report: HistGradientBoosting

	Precision	**Recall**	**F1-Score**	**Support**
Atrial Fibrillation	0.99	0.41	0.53	24
Coronary Atherosclerosis	0.98	0.04	0.04	90
Hypertension	1.00	0.01	0.02	2,494
Other	0.95	1.00	0.96	3,126
Accuracy			0.97	5,734
Macro Avg	0.98	0.35	0.39	5,734
Weighted Avg	0.94	0.94	0.91	5,734

A.6 LightGBM

Table A.6: Classification Report: LightGBM

	Precision	**Recall**	**F1-Score**	**Support**
Atrial Fibrillation	0.00	0.00	0.00	24
Coronary Atherosclerosis	0.00	0.00	0.00	90
Hypertension	0.09	0.03	0.02	2,494
Other	0.95	0.96	0.97	3,126
Accuracy			0.93	5,734
Macro Avg	0.25	0.25	0.25	5,734
Weighted Avg	0.88	0.92	0.90	5,734