CIS 5200 Fall 2025, Project Report

Early and Late Fusion Methods for Cardiovascular Disease Prediction from Longitudinal EHR and Genetic Data

Team Name: coHERent

Group Members:

- Ananya Rajagopalan; Email: ananya.rajagopalan@pennmedicine.upenn.edu
- Nia Abdurezak; Email: nia.abdurezak@pennmedicine.upenn.edu
- Sophie Kearney; Email: sophie.kearney@pennmedicine.upenn.edu

Abstract

Cardiovascular disease (CVD) is a leading cause of morbidity worldwide, and early identification and accurate risk prediction can lead to better outcomes through preventative interventions. Traditional risk assessment methods rely on a small set of baseline measurements, which may not capture longitudinal trends in electronic health record (EHR) data and the contribution of genetic risk. In this study, we develop machine learning methods and fusion techniques to predict a patient's future CVD diagnosis from longitudinal EHR data across a four year observation window and static genetic data. To process the temporal EHR data, we calculated summary statistics and a missingness indicator for each feature by aggregating across the entire observation window and partitioning into six-month slices. We incorporated genetic information with early fusion, where genotype dosage is appended to EHR records, and late fusion, where polygenic and EHR-derived risk scores are combined using a meta-model. We tested four models across all data representations: Logistic Regression, Random Forest, XGBoost and TabPFN. Across all models, integrating genetic data provides significant improvement in CVD prediction, averaging +0.29 in F1 scores compared to EHR-only baselines. TabPFN has the strongest performance in EHR-only data, but adding genetic data with early fusion leads to comparative performance across TabPFN, Random Forest, and XGBoost. We selected Random Forest and TabPFN for the late fusion analysis and maintained a similar performance gain compared to EHR data alone, although less improvement than with early fusion. These findings indicate the importance of incorporating longitudinal and genetic data for risk prediction models in a multimodal clinical setting, and thoughtful selection of a representation method when combining data modalities.

1 Motivation

CVD continues to be a leading cause of morbidity, with a global disease burden that has only increased since 1990 [11]. A variety of predictive models have been developed and applied in order to better predict and prevent CVD, but conventional models stick to well-established, clinical risk factors that are limited in generalizing to all patients [14]. Moving beyond traditional risk factors and incorporating multiple data types (e.g., genetic data, Electronic Health Records (EHR), imaging, and unstructured clinical text) for a single patient is instrumental to disease prediction tasks. Machine learning (ML) is particularly valuable for this because of models' ability to model complex, non-linear relationships among large-scale data, and doesn't always require pre-engineered features (rather, the model can *learn* the most discriminative features). Despite the promise of using ML with multimodal biomedical data, integrating multimodal data has methodological challenges. It is difficult to ensure that all data types are effectively represented when combining data modalities with different structures, resolutions, and scales. In this project, we explore machine learning techniques to integrate clinical and genetic data for cardiovascular disease prediction.

2 Related Work

Machine learning holds great promise to realizing precision medicine by personalizing medical processes (e.g., generating disease diagnoses) to patients' health. Modeling biomedical data longitudinally is a key to this goal, and multimodal information from electronic health records (EHRs) can more comprehensively capture patient health. A 2022 scoping review [8] noted that few studies at that time compared multimodal versus unimodal approaches or alternative data fusion strategies, and had limited interpretation of the model outputs for clinical audiences. The most relevant prior research to our investigation is Zhao et al. (2019) who used longitudinal EHR and genetic data from Vanderbilt University Medical Center in predicting 10-year CVD events[14]. While they demonstrated the strength of Gradient Tree Boosting in modeling these data, they did not compare fusion approaches for incorporating genetic data (they only used late-fusion), and provided no analysis of features/model output for deep learning models (CNN and LSTM) due to their black box nature. However, deep learning approaches of modeling EHR data are still valuable because they learn features automatically from raw, unstructured data, and can be designed to handle sequential data. In particular, transformers (an advancement over RNN-based approaches) have been applied to longitudinal health modeling and recent work by Niu et al. (2024) [9] and Ding et al. (2024) [3] demonstrates that transformer-based approaches are valuable for disease diagnosis tasks using clinical data. However, there is limited exploration of using transformers to model multimodal EHRs with genetic data [10].

Building on these findings, we prioritize using interpretable model architectures in this work, benchmarking our multimodal versus unimodal models with different fusion approaches to discern gains in predictive performance, and conducting feature analyses on all models. **We contribute to this problem in the following novel ways: 1) evaluating early and late-stage data fusion approaches for multimodal CVD prediction, and 2) exploring ML methods (including transformers) to model multimodal EHR data longitudinally**.

3 Data Set

We use the Coherent Data Set in this study, which is a synthetic, multimodal clinical dataset designed to emulate realistic longitudinal patient trajectories [12] with a focus on CVD. The dataset incorporates a variety of patient-level data, such as genetic data, MRI images, clinical notes, and longitudinal routine clinical observations. We focus on a subset of this data, including structured EHR clinical observations and genotype dosage data. EHR clinical observations include standard biometric and laboratory information taken at a hospital visit, such as height, BMI, lipid panels, and blood glucose levels, which are commonly used by physicians to aid CVD risk assessment. The observation table contains 1.8 million entries, with one row per patient per observation per date. Genetic data is provided as genotype dosage values for 161 selected variants, where a value of 0, 1, or 2 indicates the number of copies of the alternate allele. We converted genetic data from a long, variant-level table into a wide patient-level matrix, with each column representing a genotype dosage.

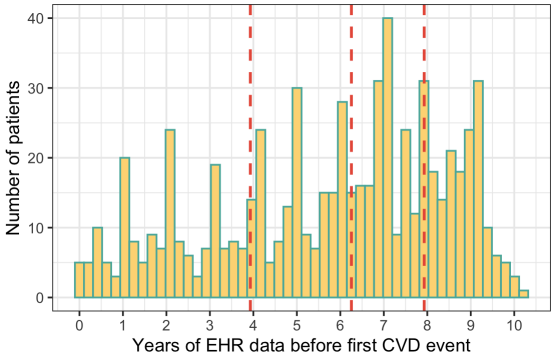

Figure 1: **Distribution of EHR history prior to first CVD event** for each CVD-positive patient with a CVD event occurring after their first EHR observation. Vertical lines indicate quartiles of the distribution.

Clinical observation data are available for 3,539 patients. Figure 1 demonstrates the distribution of the maximum number of days between a patient's earliest available EHR observation and their first CVD diagnosis. To include as many patients as possible while preserving enough training data, we defined the observation window W_{obs} as the first four years of each patient's EHR record. The first quartile of this distribution as 3.93 years, showing that around 25% of CVD-positive patients had fewer than four years of EHR-history before their first documented CVD event. The remaining data beyond four years is defined as the prediction window, W_{pred}.

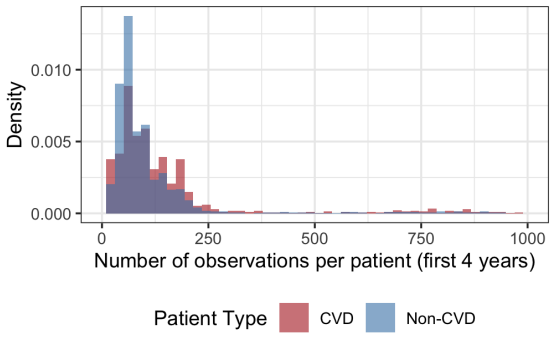

Figure 2: **Distribution of EHR observations in CVD and Non-CVD patients** within the first four years of their records.

After applying the observation window restriction to CVD patients, we obtain 3,447 patients, with 1,217 CVD and 2,230 non-CVD. Among these, 809 patients (532 CVD and 277 non-CVD) additionally had matched genetic data available and were included in the multimodal analyses. Within the four year observation window for EHR data, we examined the distribution of the number of observations each patient had for CVD and Non-CVD classes (Figure 2). We observe class imbalance in the EHR observations (63% Non-CVD vs 37% CVD), but the distributions of total observations within W_{obs} are similar between the two classes. Despite the class imbalance, the amount of data per-patient in each class is similarly distributed, so we chose not to resample and balance the dataset further.

Of the 167 unique observation types in the EHR data, we selected 53 for this study based on their biological and clinical relevance to CVD as an outcome and their likelihood of being obtained at a routine patient visit, determined through a literature search. For downstream processing, we pivoted the EHR observations table wider to create a table with one row per patient per visit and all 53 features as columns.

4 Problem Formulation

We frame CVD prediction as a supervised binary classification task at the patient level from a prognosis perspective. For each patient $p_i \in P$, we define an observation window W_{obs} containing the first four years of data from baseline and a prediction window W_{pred} containing the remaining observations. The label y is $y_i = 1$ if the patient shows any CVD in W_{pred} and $y_i = 0$ otherwise. We chose the four year threshold based on the distribution of the time to first CVD event to balance cohort size and preserve as much data as possible.

CVD is defined with the SNOMED CT identifier 49601007, "Disorder of cardiovascular system." We identified CVD events using all children of CVD in the SNOMED CT ontology using the SNOMED CT identifiers provided in the dataset for patient conditions. The labels for y are therefore a composite outcome encompassing diverse CVD outcomes.

We identified three input X formulations that map clinical data to y labels for each patient:

- **American College of Cardiology and American Heart Association Pooled Cohort Risk Equation (ACC/AHA) [5]**: Calculated with pre-visit lab values, demographic data, and smoking status from the baseline W_{obs} value.
- **Aggregated** W_{obs}: For each patient, we summarized the observations in W_{obs} into a single vector per patient (minimum, maximum, standard deviation, and median). A missingness indicator is added for each feature. This yielded a 2D matrix $X_{\text{agg}} \in \mathbb{R}^{n \times 5d}$, where $5d$ represents the four summary statistics plus one missingness indicator for each of the d features.
- **Temporal** W_{obs}: We partitioned the four year observation window W_{obs} is into eight non-overlapping six-month slices. For each slice, the same summary statistics (minimum, maximum, standard deviation, and median) plus a missingness indicator represent all observations in the slice, yielding $x_{\text{temp}_i} \in \mathbb{R}^{8 \times 5d}$ for each patient. For all patients, $X_{\text{temp}} \in \mathbb{R}^{n \times 8 \times 5d}$, which is a 3D matrix. For use in traditional machine learning models, we will concatenate each visit across time to form a flattened vector for each patient and a 2D matrix where $X'_{\text{temp}} \in \mathbb{R}^{n \times (8 \cdot 5d)}$

Not every patient has all 53 types of clinical observations at every visit. For the aggregated and temporal W_{obs}, if a patient did not have a clinical observation over the four year window, that value was imputed with the population mean for that feature. In the temporal W_{obs}, if a patient is missing a feature for one six-month slice, but it has been observed in another visit in the four-year window, the feature is imputed

with the closest timepoint. Some patients may not have a visit at all within the six-month slice, and the entire slice is imputed from the closest observed six-month slice. Missingness indicators for each feature record if the value has been observed or imputed.

5 Methods

W_{obs} contains multiple modalities that differ in structure and resolution. Clinical EHR observations (laboratory results, procedures, vitals) are collected across multiple timepoints, but genetic data are static. To understand the impact of temporal granularity of the EHR data, we compared aggregation over W_{obs}, X_{agg}, to a temporally structured representation across 6-month slices, X_{temp}. When incorporating genetic data, we tested two fusion strategies, early and late fusion, to assess wether variant-level information provides predictive signal beyond polygenic risk scores. Each model will be performing a binary classification task to predict CVD vs. non-CVD.

5.1 American College of Cardiology and American Heart Association Pooled Cohort Risk Equation as a Baseline

The American College of Cardiology and American Heart Association Pooled Cohort Risk Equation (ACC/AHA) is an algorithm commonly used by healthcare providers to estimate 10-year CVD risk using the variables age, sex, race, systolic blood pressure, smoking status, total cholesterol, HDL-C, and diabetes status [5]. We chose the ACC/AHA algorithm as our baseline due to its simplicity and direct clinical relevance. Rather than directly applying the equation, we used the same features to train a logistic regression model on our dataset using the first timepoint of each patient and pre-visit features.

5.2 Model Architecture and Selection

To perform risk prediction, we tested four models on X_{agg}, X_{temp}, X_{agg} + genetic data, and X_{temp} + genetic data to predict CVD: logistic regression (LR), random forest (RF), XGBoost, and TabPFN. We tested four models (Logistic Regression, Random forest, TabPFN and XGBoost) across four data types (both aggregated and temporal data types, of EHR only data and early fusion of EHR and genetic data), for a total of 16 combinations of models and data. We trained all models with an 80/20 test-train split and a fixed seed of 42 for reproducibility. For each data representation, we optimize hyperparameters within the training set using 10-fold cross validation with Optuna (F1 as the primary objective to balance precision and recall in the context of class imbalance). To prioritize interpretability, we used Shapley Additive exPlanations (SHAP) values to identify the most important features for CVD prediction across all models.

Logistic regression provides an interpretable baseline with an L_2 regularization penalty to reduce overfitting. We tuned the regularization parameter C to adjust the strength of the L_2 penalty and balance generalization and interpretation. Logistic regression minimized binary cross-entropy loss to optimize the separation of each class.

We chose Random Forest because it can model non-linear relationships that are often present in high-dimensional clinical data, and helps to improve generalization (reduced variance) as an ensemble method with bootstrapped samples. We tuned the number of trees and maximum tree depth, which jointly create a larger ensemble, to balance the complexity of the model with overfitting. We used the Gini impurity metric to decide each split.

We chose XGBoost as a high-performance gradient boosting algorithm that can learn high-dimensional feature interactions among data. We tuned XGBoost over a broader hyperparameter space including the number of trees, learning rate, maximum depth, subsampling ratio, column subsampling, gamma, and minimum child weight. These parameters regulate factors like model complexity, feature sampling, and the minimum gain required for additional tree splits, allowing the model to adapt to the high dimensionality and heterogeneity of EHR and genetic data. XGBoost optimized a regularized log-loss objective through gradient boosting.

Lastly, we included TabPFN as a tabular foundation model alternative to the traditional machine learning methods used in this study. TabPFN is a prior-data fitted network designed for tabular prediction tasks that

performs inference in one pass without model fitting by leveraging in-context pre-training on large synthetic datasets. Therefore, it does not require hyperparameter tuning. This allowed us to evaluate if a pre-trained attention-based architecture could perform similarly to traditional supervised learning models.

5.3 Genetic Data Processing and Representation

Given our primary objective of evaluating whether the representation of genetic data influences its predictive power when incorporated alongside EHR data, we evaluated two fusion methods: late and early fusion. Each patient-specific genotype profile is composed of 161 single nucleotide polymorphisms (SNPs) that are associated with CVD (See [12], Supplementary Materials Table 1). We constructed two distinct representations of these SNPs for our fusion methods.

For the late fusion method, we built a patient specific polygenic risk score (PRS), which is a standard measure in the field of genetics to describe an individual's genetic risk for a given disease (in this case CVD). PRS for CVD are known to accurately stratify risk, and perform well (cite the AUCs from relevant studies). Traditionally, it is calculated as a linear model by taking the summation of all genetic risk alleles for a particular person, with each allele weighted by an effect size obtained from a genome wide association study (GWAS). The PRS represents a static score that can indicate an individual's relative risk for a condition, and since it is calculated with a model itself, we append this to each patient's EHR data (as an additional column). We followed standard PRS construction methods for this task, using effect sizes from a previously published CVD GWAS ([4]). From the GWAS, 111 SNPs had effect sizes to pull from the GWAS, so we only used this set (we tried exploring other GWAS to see if there were more matches, but there were comparable levels of missingness). Out of the 111 SNPs, we employed the shrinkage strategy of different P-value selection thresholds as inclusion criteria for SNPs into the score[2]. In particular, we experimented with different p-value thresholds: ≤ 0.05 (49 SNPs), $1e^{-5}$ (27 SNPs) and $5e^{-8}$ (24 SNPs). Interestingly, the different p-value thresholds did not yield significantly different performance results, so we used the middle threshold ($p \leq 1e^{-5}$) for incorporation into our ensemble.

In the early fusion approach, we use genotype dosage (vector embeddings of 0, 1, or 2) representing how many copies of the alternate allele a patient has at a given locus. Because the SNPs are inherently the key features that indicate genetic risk, we directly embed the genetic dosage of these SNPs. We hypothesized that by using this simple encoding, the model would learn the most predictive variants (as individual features) of CVD, as opposed to using the importance indicated by GWAS effect sizes.

5.4 Application of Models to Each Data Representation

All models were tested first on the EHR modality alone for both the aggregated and temporal representations. This allowed us to assess the predictive value of temporal granularity in EHR data independent of genetic risk. We then tested two methods of integrating the temporal EHR modality with the static genetic modality:

- **Early fusion:** We incorporated genetic data directly into the existing X matrices prior to model training. For both X_{agg} and X'_{temp}, we appended static genetic features encoded as genotype dosage for each variant to each patient's aggregated feature vector, $x_i = [\text{EHR}_i, \text{genetic data}_i]$ with $X_{\text{agg}} \in \mathbb{R}^{n \times 5d}$ and $X'_{\text{temp}} \in \mathbb{R}^{n \times (t \cdot 5d)}$. $5d$ represents the four summary statistics (minimum, maximum, median, and standard deviation) and a missingness indicator used to represent d features. We tested all models on both X_{agg} and $X\prime_{\text{temp}}$ with concatenated genetic features.

- **Late fusion:** We calculated genetic risk of CVD as polygenic risk scores for each patient from the genetic data using established effect sizes and curated variant lists from the literature. We derived the EHR-based CVD risk scores from the logits of the best-performing EHR-only model. We then used the paired EHR and genetic risk scores as input features to train a logistic regression meta-model for the final prediction.

6 Experiments and Results

As a baseline, we used ACC/AHA features (their natural log, as performed in the original equation) to calculate per-patient CVD risk scores, ASCVD, using a logistic regression model. The model had an AUC of 0.69. The coefficients in this run showed a negative correlation with Systolic BP at a value of -1.088, and a positive correlation with HDL-C values at 4.6, which are consistent with the known effects of these CVD biomarkers.

For EHR-only data all models struggled in the binary classification task, across both X_{agg} and X_{temp}, with F1 scores around 0.50. The addition of genetic data, through both early and late fusion, greatly improves predictive power as seen in Figure 3. For example, logistic regression showed a significant increase in F1 scores, with +0.37 and +0.25 for X_{agg} and X_{temp} respectively.

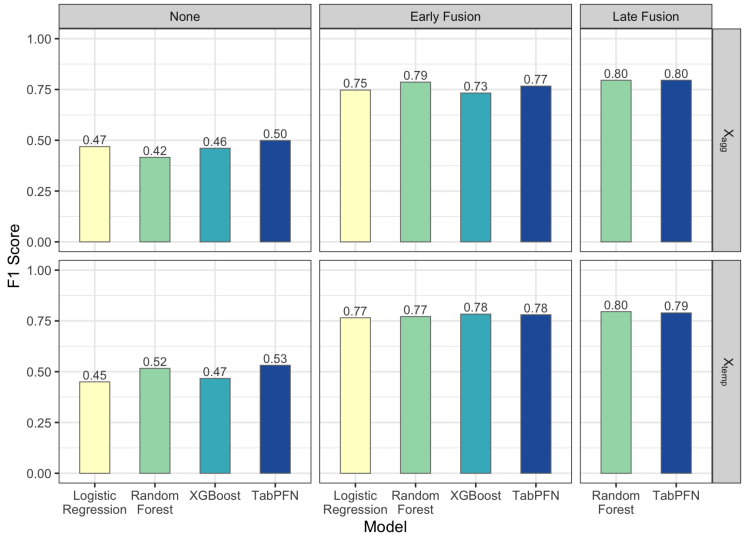

Figure 3: **Comparison of F1 scores** of all models across genetic data fusion strategies and EHR temporal data representation strategies.

For both aggregated and temporal data types, the addition of genetic data with early fusion showed an increase in F1, precision, recall and AUPRC across all models, although some models had a slight decrease in both balanced accuracy and AUROC. Random forest decreased in AUROC for X_{agg} with the addition of genetic data for early fusion from 0.75 to 0.62 as well as a decrease with XGBoost in the same data type from 0.75 to 0.68.

In early fusion experiments, logistic regression performed slightly worse with F1 scores of 0.75 and 0.77 for X_{agg} and X_{temp}, compared to the top performing model TabPFN with F1 scores of 0.77 and 0.78 for X_{agg} and X_{temp}. TabPFN was the best performer in the EHR-only setting, but here the performance is comparable across models: for X_{temp}, XGBoost matched TabPFN in F1 and AUPRC while TabPFN kept the highest precision, and for X_{agg}, Random Forest achieved the best F1 score. We see better classification of CVD cases in early fusion compared to EHR-only data, but more misclassified Non-CVD cases, reflected by the decrease in balanced accuracy for Logistic Regression, XGBoost and Random forest with X_{agg}.

6.1 Late Fusion

We chose TabPFN as the top-performing model on the EHR-only data across Balanced Accuracy, AUROC and F1 score to create the EHR-derived risk score for late fusion. We used normalized PRS scores to represent the input feature for genetic data. We used logistic regression as a meta-model to predict on the two risk scores from both EHR and genetic data. After observing that the class imbalance may be contributing to an overestimate of logistic regression predicting CVD, we added an additional test of EHR-only data represented by the second best performing model, Random Forest. This additional experiment had similar performance to the TabPFN-derived EHR risk scores, which suggests that the biological predictors for CVD did not translate well across separate transformations of both EHR and

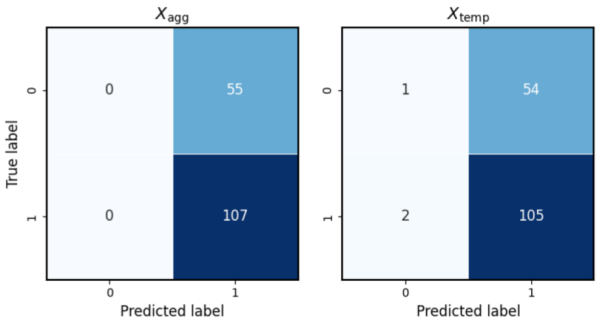

Figure 4: **Confusion matrices for the Random Forest model on Late Fusion** of genetic data to aggregated and temporal EHR data. The model predicts CVD for almost all patients, with most errors being false positives.

Genetic Fusion	Data Type	Model	F1	Precision	Recall	Balanced Acc.	AUROC	AUPRC
None	ACC/AHA Feat.	ASCVD	0.39	0.59	0.29	0.59	0.69	0.54
	X_{agg}	LR	0.47	0.56	0.40	0.62	0.73	0.59
		RF	0.42	0.60	0.32	0.60	0.75	0.61
		XGB	0.46	0.60	0.37	0.62	0.75	0.59
		tabPFN	**0.50**	**0.62**	**0.42**	**0.64**	**0.76**	**0.64**
	X_{temp}	LR	0.45	0.58	0.37	0.61	0.72	0.57
		RF	0.52	0.60	0.45	0.64	0.75	0.59
		XGB	0.47	0.58	0.39	0.62	0.74	0.58
		tabPFN	**0.53**	**0.62**	**0.47**	**0.65**	**0.76**	**0.61**
Early fusion	X_{agg}	LR	0.75	0.64	0.90	0.46	0.53	0.70
		RF	**0.79**	0.66	**0.96**	0.51	0.62	0.81
		XGB	0.73	0.65	0.83	0.49	0.68	0.84
		tabPFN	0.77	**0.69**	0.86	**0.56**	**0.72**	**0.86**
	X_{temp}	LR	0.77	0.68	0.87	0.54	0.57	0.72
		RF	0.77	0.67	**0.92**	0.51	0.66	0.83
		XGB	**0.78**	**0.70**	0.90	**0.57**	0.70	**0.85**
		tabPFN	**0.78**	0.69	0.90	0.56	**0.71**	**0.85**
Late fusion	X_{agg}	RF	**0.80**	**0.66**	**1.00**	**0.50**	**0.57**	**0.74**
		tabPFN	**0.80**	**0.66**	**1.00**	**0.50**	0.50	0.72
	X_{temp}	RF	**0.80**	**0.66**	**1.00**	**0.50**	**0.57**	**0.73**
		tabPFN	0.79	**0.66**	0.98	**0.50**	0.51	0.72

Table 1: **Model performance across experiments** with aggregated and temporal EHR data, with and without fusion of genetic data.

genetic data in late fusion (compared to early fusion). The over-prediction the CVD class continued when using the random forest model for EHR-derived risk scores, as seen in Figure 4.

6.2 Feature Importance

Given our interest in model interpretability, we wanted to explore how feature importance changed across the different data modalities used (when using EHR only data to combining EHR + genetic data modalities), as well as across different models for a given data type (e.g., X_{temp}) (see Figure 5). For the latter comparison, this would indicate to what extent certain features are truly driving CVD risk.

7 Conclusion and Discussion

7.1 Key Findings: Genetic data is valuable, but representation method matters

Our primary contribution in this work is exploring the advantage of incorporating genetic data alongside traditional EHR data for disease risk prediction. In particular, we built on previous work by comparing different representations of genetic data (early fusion versus late fusion) and whether that influences overall model performance.

Our first key finding is that incorporating genetic data is valuable for CVD risk prediction, beyond just EHR data: across all model architectures, adding genetic data alongside EHR data (regardless of the fusion method) improved model performance across F1, Precision, Recall, and AUPRC compared to EHR data alone. Additionally, TabPFN outperformed other architectures across all metrics when modeling EHR data only, and when adding genetic data using early fusion, model performance was comparable across RF, XGBoost and TabPFN.

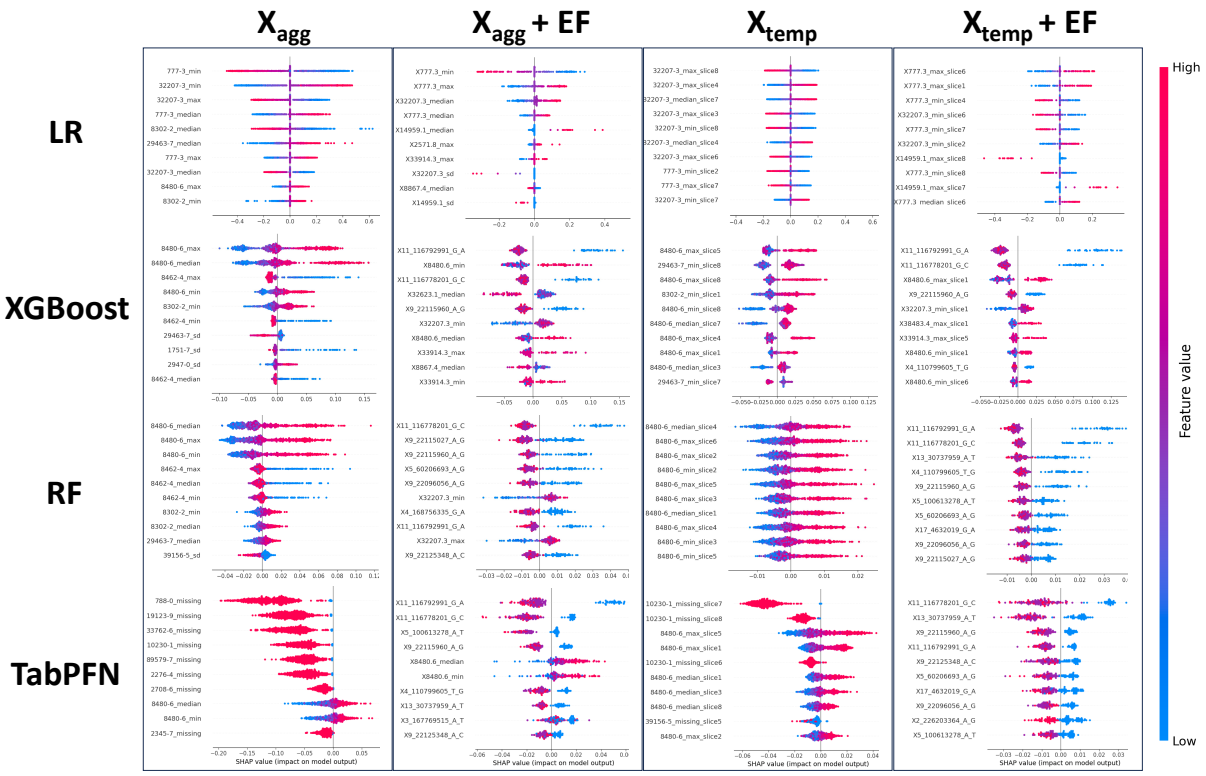

Figure 5: **Comparison of top 10 feature importance by SHAP across all model architectures and data types.** X axis of each subplot contains the SHAP value per feature, indicating the magnitude of a feature's contribution to a given model's output. Y axis indicates the value of a feature itself. EF = **E**arly **F**usion of genetic data with EHR data. Note: for the columns with EF data, there is an 'X' character appended to certain EHR features (LOINC standard vocabulary) and genetic variants (SNP IDs) that does not belong there. E.g., 777.3_min is the same feature across both LR plots for X_{agg} and $X_{agg} + EF$. The Supplementary folder (Github) has a table containing full LOINC to feature mappings.

Our second key finding was that using early fusion to incorporate genetic data led to better model performance compared to late fusion. This connects to our initial hypothesis of the importance of data representations, in particular, that incorporating the genetic data using early fusion would provide richer information for the model (which influences model performance). This finding represents the primary novelty of this work, since the relevant papers we found [14] did not explore the value of different fusion methods. Our results suggest that a single PRS may be overly restrictive for a model, particularly in cases where the PRS constructed does not have a particularly strong predictive performance. This is a common issue for PRS in many complex diseases, where differences in ancestral composition of patient populations or underpowered GWAS [13] can undermine the predictive value of a scalar score. Additionally, given that late fusion involves a meta-model (model of models) approach, we may need additional calibration beyond hyperparameter optimization to ensure that predictive error does not propagate. Alternatively, the method we used for early fusion (using variant-level information by constructing additive genotype dosage embeddings) provides more information from the model to learn from. The model can essentially decide for itself which variants are the most predictive of disease in the given patient population, leading to a more robust model.

7.2 Feature Analysis: Exploring Clinical Relevance

To qualitatively understand whether our model outputs were clinically relevant to prognosticating CVD risk, we conducted a feature analysis using SHAP. Figure 5 contains a comprehensive visualization of the top 10

features for each model type and dataset. Here, we discuss its most salient findings and interpretations.

We observed the following general trends when looking at the feature importance across models: 1) the top features are clinically relevant to CVD, indicating that the models learned genuine predictors of disease risk, and 2) there is a subset of clinical and genetic features that persist in their high importance among model types, which supports their relevance. EHR features are encoded by LOINC terms, which we queried to understand the clinical concepts behind each of the codes (see the supplementary folder on Github for a table containing detailed code to feature mappings). The following EHR features maintained high importance across multiple, if not all, model architectures and data inputs: 777-3 (Platelets in Blood by Automated count), 8480-6 (Systolic Blood Pressure), 32207-3 (Platelet distribution width in Blood by Automated count), 29463-7 (Body Weight), 8302-2 (Body Height), and 8462-4 (Diastolic Blood Pressure). Evidently, these features both align with and go beyond those used in the ACC/AHA baseline equation, particularly with molecular biomarkers of CVD. Other clinical features that don't persist across multiple models are still physiologically relevant to CVD risk (e.g., 39156-5: Body Mass Index, 2571-8:Triglycerides, 8867-4:Heart rate), highlighting the benefit of using several model architectures that can model nuanced relationships within data.

With regards to the addition of genetic data through early fusion, we see a similar pattern of certain genetic variants (as features) being consistently among the most important across multiple model types, and others coming up more sporadically in the top 10 set. Three genetic variants that persisted across RF, XGBoost and TabPFN models with both X_{agg} and X_{temp} data were 11_116792991_G_A (*rs662799*), 11_116778201_G_C (*rs964184*), and 9_22115960_A_G (*rs2383207*), which all have previously reported physiological relevance to CVD and were in the top 25 SNPs with the smallest p-values from the previously reported GWAS [4]. *rs662799*, located in the *APOA5* gene has been linked to higher levels of plasma triglycerides and is consequently associated with an increased CVD risk [6]. *rs964184* in the *ZPR1* gene has been associated with blood lipid levels, as well as with the risk of myocardial infarction among high-risk cardiovascular patients [1]. Thirdly, *rs2383207* in chromosome 9 has been previously associated with coronary artery disease [7].

The addition of genetic data to EHR data resulted in a prioritization of a mix of both feature types across all model architectures, interestingly with the exception of logistic regression. Additionally, while TabPFN was the best at predicting CVD from EHR data alone, its feature importance analysis revealed some ambiguity by predominantly having missingness indicator variables as the biggest contributors to the output. This is in direct contrast to the trends in feature importance across all other models. Perhaps even more surprisingly, this trend disappears when genetic data is added to TabPFN, perhaps by forcing the model to tune more to biological signal among the positive cases. In its current state, TabPFN appears to have limited clinical utility compared to other models with EHR data alone, despite its strong prediction of CVD risk.

7.3 Lessons Learned

During the model development process, we decided to standardize our experimental setup (i.e., same hyperparameter tuning in the number of folds and library used, consistent random seeds and stratifying on the target variable, same train-test split) to enable a fair and robust comparison of model performance. We didn't consider doing this when we initially started the project, but realized later the importance of this step when evaluating many different models and datasets.

When constructing the PRS to be used in late fusion of genetic data with EHR data, we explored different p-value thresholds (see Section 5) to ensure that any highly predictive variants would not be left out of the score construction process. We learned that the different p-value thresholds did not make a substantial impact on the overall performance of the PRS, which we realized was biologically sensible since the *most* significant variants were ultimately the most important (predictive) of CVD. Given that the overall predictive power of the PRS for CVD was still mediocre compared to previously published CVD PRS, we think that there may be limitations coming from the fact that this is a synthetic dataset, so the signal of these variants could be diluted (even though they were pre-selected for their association with CVD risk).

Our third learning was with regards to considering what cases versus control 'profiles' may look like for certain features. More specifically, we observed across all of our models that when we added in genetic data along with EHR data, we observed better performance overall, but when examining the confusion matrices

(see 4) for these models, we noticed that the models were predicting many more cases than controls using the genetic + EHR data (this is also suggested by the decreased balanced accuracy and AUROC, despite increases across all other metrics). The increase in F1, Precision, Recall, and AUPRC suggests that the model gets better at identifying positive cases when we add genetic data, but at the cost of overall model performance in distinguishing between cases and controls. This was quite perplexing to us at first: we experimented with optimizing for a different parameter (not F1) during hyperparameter optimization which did not help the skew towards predicting positive cases. We then considered how it is possible for the model to do better at predicting cases at the expense of overall discrimination against controls. We thought it may be because the addition of all variant information as genotype dosage vectors has a clear trend for the model to determine cases (a certain subset of variants are very strongly associated with CVD), but non-disease associated variants may have natural variation in their presence among patients, which makes it more difficult for the model to learn a very clear 'profile' of the controls, as it does for the cases. As a result, even though the overall model performance improves, there is a less generalizable representation/separation of cases and controls when we add the variant-level genetic data. This could be further investigated with additional genetic data that has a similar class balance of cases versus controls to the original dataset, which was not the case in this work.

7.4 Future Directions: Additional Modalities and Architectures

There are many opportunities for future research and extensions to our project. First, we could incorporate additional data modalities that are included in the Coherent data set, such as imaging (DICOM files of MRI scans) and text (simplified clinical notes, 100+ per patient). We de-prioritized including these additional modalities for the scope of this project, since we wanted to focus more on the inclusion of genetic data and experimenting with various representations of it. For the use of text data in particular, we discussed using the Open AI API to generate embeddings of each clinical note and using these vectors as additional features in our models. Clinical notes from the first 4 years of a patient history (prior to a CVD diagnosis) could be used here. The second way of extending our work is by experimenting with the performance of different models that are designed for modeling sequential data: RNNs and LSTMs. We could also explore the use of these deep learning models for representing genetic data differently. We did not use these models because of their limited interpretability and worse performance compared to RF and XGBoost models in prior work. However, given that TabPFN performed strongly with our dataset but had some surprising results during feature importance analysis (see 7.2), other deep learning approaches tailored for sequential data are worth exploring.

Code Availability

All code used in this project is available at `https://github.com/sophie-kearney/CVD_multimodal_prediction`.

Acknowledgments

We thank Prof. Lyle Ungar, our recitation TA Allain Welliver, and our final project TA Lulu Liu for their support and advice throughout this course. Additionally, we thank the authors of the Coherent synthetic dataset for generating such an accessible and unique (multimodal) resource for us to explore in this project.

References

[1] Juan F. Alcala-Diaz et al. "A Gene Variation at the ZPR1 Locus (rs964184) Interacts With the Type of Diet to Modulate Postprandial Triglycerides in Patients With Coronary Artery Disease: From the Coronary Diet Intervention With Olive Oil and Cardiovascular Prevention Study". In: *Front. Nutr.* 9 (June 2022), p. 885256. ISSN: 2296-861X. DOI: 10.3389/fnut.2022.885256. URL: https://www.frontiersin.org/articles/10.3389/fnut.2022.885256/full (visited on 11/26/2025).

[2] Shing Wan Choi et al. "Tutorial: a guide to performing polygenic risk score analyses". en. In: *Nat Protoc* 15.9 (Sept. 2020), pp. 2759–2772. ISSN: 1754-2189, 1750-2799. DOI: 10.1038/s41596-020-0353-1. URL: https://www.nature.com/articles/s41596-020-0353-1 (visited on 12/02/2025).

[3] Jun-En Ding et al. *Large Language Multimodal Models for 5-Year Chronic Disease Cohort Prediction Using EHR Data*. Version Number: 2. 2024. DOI: 10.48550/ARXIV.2403.04785. URL: https://arxiv.org/abs/2403.04785 (visited on 10/27/2025).

[4] Handan Melike Dönertaş et al. "Common genetic associations between age-related diseases". en. In: *Nat Aging* 1.4 (Apr. 2021), pp. 400–412. ISSN: 2662-8465. DOI: 10.1038/s43587-021-00051-5. URL: https://www.nature.com/articles/s43587-021-00051-5 (visited on 11/10/2025).

[5] David C. Goff et al. "2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines". en. In: *Circulation* 129.25_suppl_2 (June 2014). ISSN: 0009-7322, 1524-4539. DOI: 10.1161/01.cir.0000437741.48606.98. URL: https://www.ahajournals.org/doi/10.1161/01.cir.0000437741.48606.98 (visited on 11/10/2025).

[6] Jerry Jacob et al. "Apolipoprotein A5 gene polymorphism (rs662799) and cardiovascular disease in end-stage kidney disease patients". en. In: *BMC Nephrol* 23.1 (Sept. 2022), p. 307. ISSN: 1471-2369. DOI: 10.1186/s12882-022-02925-1. URL: https://bmcnephrol.biomedcentral.com/articles/10.1186/s12882-022-02925-1 (visited on 11/26/2025).

[7] Shiridhar Kashyap et al. "The association of polymorphic variants, rs2267788, rs1333049 and rs2383207 with coronary artery disease, its severity and presentation in North Indian population". en. In: *Gene* 648 (Mar. 2018), pp. 89–96. ISSN: 03781119. DOI: 10.1016/j.gene.2018.01.021. URL: https://linkinghub.elsevier.com/retrieve/pii/S0378111918300283 (visited on 11/26/2025).

[8] Adrienne Kline et al. "Multimodal machine learning in precision health: A scoping review". en. In: *npj Digit. Med.* 5.1 (Nov. 2022), p. 171. ISSN: 2398-6352. DOI: 10.1038/s41746-022-00712-8. URL: https://www.nature.com/articles/s41746-022-00712-8 (visited on 10/27/2025).

[9] Shuai Niu et al. "EHR-KnowGen: Knowledge-enhanced multimodal learning for disease diagnosis generation". en. In: *Information Fusion* 102 (Feb. 2024), p. 102069. ISSN: 15662535. DOI: 10.1016/j.inffus.2023.102069. URL: https://linkinghub.elsevier.com/retrieve/pii/S1566253523003858 (visited on 10/26/2025).

[10] Clauirton A. Siebra et al. "Transformers in health: a systematic review on architectures for longitudinal data analysis". en. In: *Artif Intell Rev* 57.2 (Feb. 2024), p. 32. ISSN: 1573-7462. DOI: 10.1007/s10462-023-10677-z. URL: https://link.springer.com/10.1007/s10462-023-10677-z (visited on 10/27/2025).

[11] Benjamin A. Stark et al. "Global, Regional, and National Burden of Cardiovascular Diseases and Risk Factors in 204 Countries and Territories, 1990-2023". en. In: *JACC* 86.22 (Dec. 2025), pp. 2167–2243. ISSN: 07351097. DOI: 10.1016/j.jacc.2025.08.015. URL: https://linkinghub.elsevier.com/retrieve/pii/S0735109725074285 (visited on 11/27/2025).

[12] Jason Walonoski et al. "The "Coherent Data Set": Combining Patient Data and Imaging in a Comprehensive, Synthetic Health Record". en. In: *Electronics* 11.8 (Apr. 2022), p. 1199. ISSN: 2079-9292. DOI: 10.3390/electronics11081199. URL: https://www.mdpi.com/2079-9292/11/8/1199 (visited on 10/26/2025).

[13] Ying Wang et al. "Challenges and Opportunities for Developing More Generalizable Polygenic Risk Scores". en. In: *Annu. Rev. Biomed. Data Sci.* 5.1 (Aug. 2022), pp. 293–320. ISSN: 2574-3414, 2574-3414. DOI: 10.1146/annurev-biodatasci-111721-074830. URL: https://www.annualreviews.org/doi/10.1146/annurev-biodatasci-111721-074830 (visited on 11/27/2025).

[14] Juan Zhao et al. "Learning from Longitudinal Data in Electronic Health Record and Genetic Data to Improve Cardiovascular Event Prediction". en. In: *Sci Rep* 9.1 (Jan. 2019), p. 717. ISSN: 2045-2322. DOI: 10.1038/s41598-018-36745-x. URL: https://www.nature.com/articles/s41598-018-36745-x (visited on 11/10/2025).