Journal of Science, Innovation & Social Impact
EISSN: 3105-5028 | PISSN: 3105-501X | Vol. 2, No. 1 (2026)

Article

Fairness-Aware Multimodal Fusion for Early Chronic Disease Risk Prediction: A Temporal Deep Learning Approach

Xiaotong Shi [1,*]

[1] Business Analytics, Columbia University, NY, USA
* Correspondence: Xiaotong Shi, Business Analytics, Columbia University, NY, USA

Abstract: Chronic diseases constitute a significant public health challenge, with early detection enabling effective preventive interventions. This paper introduces a fairness-aware framework integrating multimodal health data-electronic health records, medical imaging, genomics, and wearable sensors-for early chronic disease risk prediction. The approach addresses three critical challenges: cross-modal feature harmonization across heterogeneous data types, algorithmic bias mitigation through fairness-constrained learning, and temporal pattern extraction for disease progression modeling. Evaluation on diabetes, cardiovascular disease, and cancer prediction using MIMIC-IV, UK Biobank, and wearable device cohorts demonstrates superior performance (AUROC: 0.892-0.924) while maintaining demographic parity across age, sex, and racial groups, while maintaining demographic parity across age, sex, and racial groups, using each cohort's available modalities where applicable. Fairness metrics improve by 76.8% relative to baseline approaches (reducing the maximum subgroup AUROC gap) without sacrificing predictive accuracy, demonstrating that equitable healthcare AI is achievable through integrated fairness-aware design.

Keywords: multimodal health data fusion; algorithmic fairness; temporal deep learning; chronic disease prediction

Received: 21 December 2025
Revised: 10 February 2026
Accepted: 22 February 2026
Published: 27 February 2026

1. Introduction

1.1. Background and Motivation

1.1.1. Rising Burden of Chronic Diseases and the Need for Early Detection

Chronic diseases, including diabetes, cardiovascular disorders, and cancers, account for 71% of global mortality, imposing substantial economic burdens on healthcare systems [1]. In the United States, chronic disease management consumes approximately 90% of annual healthcare expenditures. The extended latency period before clinical symptom manifestation presents critical windows for intervention, as early detection and preventive measures can substantially alter disease progression trajectories.

Traditional screening approaches rely predominantly on single-modality assessments-fasting glucose for diabetes, periodic imaging for cancer-which capture only fragmentary aspects of health status. The multifactorial nature of chronic disease etiology necessitates comprehensive assessment strategies. Population studies indicate that integrating diverse data modalities can enhance risk stratification precision by 15-40% compared to conventional single-marker approaches [2].

1.1.2. Opportunities and Challenges in Multimodal Health Data Integration

The proliferation of digital health technologies has generated unprecedented volumes of heterogeneous health data. Electronic health record systems capture

longitudinal clinical observations spanning years. Medical imaging modalities provide detailed anatomical information. Genomic sequencing reveals individual genetic risk profiles. Wearable biosensors continuously monitor physiological parameters.

Different modalities exhibit disparate statistical properties, dimensional scales, temporal resolutions, and semantic representations. Medical images comprise high-dimensional spatial data. Electronic health records consist of irregular time series mixing structured measurements with unstructured narratives. Genomic data present static but extremely high-dimensional feature spaces [3]. Effective fusion architectures must accommodate these heterogeneous characteristics while extracting meaningful cross-modal relationships.

1.2. Research Gaps and Challenges

1.2.1. Limitations of Unimodal Approaches in Capturing Comprehensive Health Status

Current disease prediction methodologies predominantly operate within single data modalities, constraining representational capacity. Risk scores derived solely from electronic health records overlook imaging biomarkers that reveal early pathological changes. This fragmented analytical landscape results in incomplete risk assessment.

Recent investigations documented substantial performance gains when integrating multiple modalities. Research examining cardiovascular risk prediction found that combining electrocardiogram data with laboratory measurements improved AUROC from 0.78 to 0.86. Cancer detection studies demonstrated that fusing histopathology images with molecular profiles increased accuracy by 12-18 percentage points [4]. Despite encouraging findings, systematic frameworks for multimodal integration remain underdeveloped.

1.2.2. Algorithmic Fairness Concerns in Clinical Prediction across Diverse Populations

Machine learning models deployed in healthcare settings exhibit systematic performance disparities across demographic subgroups defined by race, ethnicity, sex, and socioeconomic status. These fairness gaps emerge through multiple mechanisms: historical underrepresentation of minority populations in training datasets, systematic differences in data quality across demographic groups, and algorithmic reliance on features correlated with protected attributes [5]. Such biases can perpetuate existing health inequities if deployed without appropriate safeguards.

1.2.3. Underutilization of Temporal Patterns in Disease Progression

Chronic disease development follows dynamic trajectories characterized by gradual accumulation of risk factors, episodic acute events, and variable progression rates. Static snapshot assessments neglect this temporal dimensionality. The timing and sequence of clinical events convey prognostic information that is not captured by simple aggregation of historical observations.

1.3. Contributions

This work advances multimodal health data analytics through three principal contributions. First, we develop a feature harmonization framework that aligns heterogeneous data modalities into unified representation spaces while preserving modality-specific characteristics. Our cross-modal alignment strategy employs contrastive learning objectives to establish semantic correspondences.

Second, we introduce a fairness-constrained learning framework that incorporates demographic-parity-style and equalized-odds-style considerations directly into the optimization objectives [6]. Unlike post-hoc bias correction methods, our approach jointly optimizes predictive accuracy and fairness throughout training.

Third, we design hierarchical temporal attention mechanisms capturing disease progression patterns at multiple timescales, enabling identification of critical time windows for intervention. We validate our framework on diabetes, cardiovascular disease, and cancer prediction tasks.

2. Related Work

2.1. Multimodal Data Fusion in Healthcare

2.1.1. Fusion Strategies: Early, Intermediate, and Late Fusion Techniques

Multimodal fusion architectures span a spectrum of integration strategies. Early fusion concatenates raw features from different modalities before applying machine learning models, enabling algorithms to learn cross-modal interactions directly. This approach provides maximal flexibility but faces challenges with dimensionality management. Late fusion maintains separate processing pipelines for each modality, combining predictions only at the final decision stage [7].

Intermediate fusion represents a middle ground, allowing modalities to undergo initial independent processing before merging at intermediate representation levels. Recent architectures employ attention mechanisms at fusion points, enabling models to weight modality contributions dynamically based on task-specific requirements.

2.1.2. Deep Learning Architectures for Cross-Modal Representation Learning

Deep learning has catalyzed advances in learning unified representations from multimodal health data. Transformer architectures handle heterogeneous medical data through specialized tokenization schemes and apply self-attention mechanisms to capture intra- and inter-modal dependencies [8].

Contrastive learning frameworks learn to project different modalities into shared embedding spaces where semantically similar instances cluster together. By maximizing agreement between corresponding samples across modalities while pushing apart unrelated instances, these approaches establish semantic correspondences [9]. Graph neural networks offer another paradigm, representing data modalities as nodes and relationships as edges in heterogeneous graph structures.

2.2. Fairness and Bias Mitigation in Clinical Prediction

2.2.1. Sources of Algorithmic Bias in Healthcare AI

Algorithmic bias in healthcare prediction originates from multiple sources throughout machine learning pipelines. Historical biases embedded in training data reflect systematic inequities in healthcare access, utilization patterns, and documentation practices. Minority populations often have sparser medical records with lower documentation quality.

Measurement bias arises from systematic differences in data collection across populations. Pulse oximetry measurements are less accurate in individuals with darker skin pigmentation. Diagnostic imaging protocols may vary across institutions serving different demographic compositions [10]. Label bias occurs when definitions of outcomes differ systematically across populations.

2.2.2. Fairness Metrics and Mitigation Strategies in Medical Machine Learning2.2.2. Fairness Metrics and Mitigation Strategies in Medical Machine Learning

Quantifying algorithmic fairness requires metrics capturing different dimensions of equitable performance. Demographic parity requires that favorable outcomes occur at equal rates across protected groups. Equalized odds require that the true-positive and false-positive rates be equal across groups [11].

Pre-processing mitigation strategies address bias by modifying training data before model development. In-processing methods incorporate fairness constraints directly into learning objectives, jointly optimizing accuracy and equitable performance. Adversarial debiasing trains fairness critics alongside prediction models [12].

2.2.3. Trade-Offs between Predictive Accuracy and Demographic Parity

Fundamental tensions exist between maximizing aggregate predictive accuracy and ensuring fairness across demographic subgroups. Pareto frontier analyses characterize achievable combinations of accuracy and fairness metrics. Research examining sepsis prediction found that modest accuracy reductions (1-2% in AUROC) enabled substantial fairness improvements [13]. Context-specific considerations shape acceptable trade-off points.

2.3. Temporal Modeling for Disease Progression

2.3.1. Time-Aware Attention Mechanisms for Longitudinal EHR Data

Electronic health records pose unique temporal modeling challenges due to irregular sampling, variable-length histories, and complex dependencies between historical events and future outcomes. Time-aware attention mechanisms address these challenges by learning to weight historical observations based on relevance for prediction.

Hierarchical temporal architectures aggregate information at multiple timescales, capturing both short-term physiological dynamics and long-term disease trajectories. Lower levels process fine-grained sequences of measurements within clinical encounters. Higher levels integrate encounter-level summaries over extended periods.

2.3.2. Transformer-Based Approaches for Sequential Clinical Event Prediction

Transformer architectures have emerged as powerful tools for modeling longitudinal clinical data, offering advantages in capturing long-range dependencies and enabling parallel processing. Masked modeling approaches pre-train transformers on large unlabeled clinical databases by predicting randomly masked observations [14]. This self-supervised learning captures general patterns of clinical evolution. Causal attention mechanisms restrict information flow to respect temporal ordering [15].

3. Methodology

3.1. Multimodal Feature Harmonization

3.1.1. Data Preprocessing and Modality-Specific Encoding for EHR, Imaging, Genomics, and Wearable Data

The multimodal feature harmonization pipeline initiates with modality-specific preprocessing tailored to statistical properties of each data type. Electronic health record processing addresses missing values through forward-filling for time-invariant attributes and temporal imputation for longitudinal measurements. Categorical variables undergo learned embeddings, mapping discrete codes to continuous representations that capture clinical relationships. Numerical measurements are standardized within physiologically plausible ranges [16].

Medical imaging data undergoes standardized preprocessing, including intensity normalization, spatial resampling to isotropic resolution, and anatomical alignment. Convolutional neural networks pre-trained on large-scale medical imaging repositories extract hierarchical visual features. Attention pooling aggregates spatial features into fixed-dimension vectors suitable for fusion with other modalities [17].

Genomic data processing addresses the extreme dimensionality of genomic data through biologically informed feature selection. Polygenic risk scores aggregate the effects of established disease-associated variants weighted by effect sizes from genome-wide association studies. Gene expression profiles are log-transformed, and batch effects are corrected using empirical Bayes methods. Wearable sensor streams require specialized temporal processing to extract meaningful features from continuous physiological monitoring. Heart rate variability metrics computed over multiple timescales characterize autonomic function.

3.1.2. Cross-Modal Alignment and Unified Representation Learning

Establishing semantic correspondences across heterogeneous modalities requires alignment strategies mapping diverse data types into shared representation spaces. Our cross-modal alignment framework employs contrastive learning objectives, bringing together representations from different modalities describing the same patient:

$$L_{align} = -\log \left(\exp \left(\text{sim} \left(z_i, z_j \right) / \tau \right) / \Sigma_k \exp \left(\text{sim} \left(z_i, z_k \right) / \tau \right) \right)$$

where z_i and z_j represent encoded features from different modalities for patient i, sim computes cosine similarity, τ controls temperature, and summation spans negative samples. Here, (i, j) denotes a positive pair from the same patient across modalities, and k indexes other patients within the same mini-batch as negatives [18].

Modality-specific encoders process each data type independently. Electronic health record encoders employ temporal convolutional networks with dilation. Imaging encoders utilize ResNet architectures adapted for 3D medical volumes. Genomic encoders implement multi-layer perceptrons with dropout regularization [19]. Cross-modal attention mechanisms enable explicit modeling of relationships between modalities. Query-key-value attention computes relevance scores between pairs of modalities, dynamically weighting contributions based on content. The unified representation combines aligned modality embeddings via a learned weighted sum [20].

3.2. Fairness-Constrained Learning Framework

3.2.1. Bias Detection through Subgroup Performance Analysis

Systematic evaluation of algorithmic bias requires a comprehensive assessment of model performance across demographic subgroups defined by protected attributes. Our bias detection protocol stratifies validation cohorts by combinations of age categories, biological sex, and self-reported race/ethnicity, computing standard performance metrics within each stratum.

Performance-disparity metrics quantify the magnitude of fairness violations. The maximum absolute difference in AUROC across demographic groups indicates the worst-case prediction quality gaps. Calibration curves stratified by demographic categories assess whether the predicted probabilities are consistent with the observed outcome frequencies (Table 1).

Table 1. Baseline model performance across demographic subgroups for diabetes prediction.

Demographic Subgroup	Sample Size	AUROC	Sensitivity	Specificity	PPV	NPV
Age 18-44, Female, White	8,742	0.867	0.813	0.842	0.391	0.971
Age 18-44, Female, Black	1,236	0.841	0.779	0.828	0.356	0.968
Age 45-64, Male, White	12,458	0.892	0.851	0.863	0.472	0.978
Age 45-64, Male, Black	2,147	0.874	0.826	0.849	0.441	0.975
Age 65+, Female, Hispanic	1,893	0.879	0.834	0.856	0.518	0.973
Age 65+, Male, Asian	967	0.883	0.842	0.861	0.487	0.976

Intersectional analysis examines combinations of protected attributes, recognizing that individuals occupying multiple marginalized categories may experience compounded disadvantages. Computing performance metrics across all combinations of demographic categories provides a comprehensive assessment of bias.

3.2.2. Fairness Regularization with Demographic-Aware Loss Functions

Integrating fairness objectives directly into model optimization addresses bias at its source. Our fairness-constrained learning framework augments prediction loss with penalty terms discouraging disparate performance:

L_total = L_pred + λ_fair × L_fairness + λ_reg × L_regularization

where L_pred represents standard cross-entropy prediction loss, L_fairness quantifies fairness violations. The fairness loss incorporates a combination of distributional-parity (demographic-parity-style-style) and error-rate-parity (equalized-odds-style) criteria:

L_fairness = Σ_g,g' |TPR_g - TPR_g'| + |FPR_g - FPR_g'| + KL (P_g || P_g')

Summing over pairs of demographic groups $g$and g', where TPR and FPR denote true and false positive rates (Table 2).

Table 2. Accuracy-fairness trade-off analysis across fairness regularization strengths.

Model Configuration	Overall AUROC	Min Subgroup AUROC	Max AUROC Gap	Equalized Odds Diff	Calibration Error
Baseline (no fairness)	0.897	0.841	0.056	0.124	0.038
Fair-Reg (λ_fair = 0.1)	0.894	0.862	0.032	0.089	0.029
Fair-Reg (λ_fair = 0.3)	0.887	0.874	0.013	0.047	0.021
Fair-Reg (λ_fair = 0.7)	0.879	0.876	0.003	0.018	0.017
Adversarial Debiasing	0.891	0.868	0.023	0.062	0.024

We further incorporate an adversarial branch that predicts protected attributes from learned representations, and apply gradient reversal to discourage encoding demographic signals. Gradient regularization techniques directly constrain the influence of protected attributes on predictions. Min Subgroup AUROC denotes the minimum AUROC across the predefined demographic subgroups used in the subgroup analysis.

3.2.3. Balancing Sensitivity, Specificity, and Equitable Performance across Populations

Clinical decision support systems must maintain acceptable performance across multiple criteria: overall predictive accuracy, sensitivity for detecting true positives, specificity for avoiding false positives, and equitable performance across demographic groups. Achieving clinically acceptable trade-offs requires explicit multi-objective optimization strategies.

Pareto optimization identifies models on the accuracy-fairness frontier. Threshold optimization provides another mechanism for balancing objectives post-training. Rather than applying uniform decision thresholds, group-specific thresholds can be selected to equalize desired performance metrics. Uncertainty-aware prediction generates confidence estimates alongside risk scores, enabling deployment strategies to defer high-uncertainty cases to human review.

3.3. Temporal Pattern Extraction and Risk Stratification

3.3.1. Hierarchical Time-Aware Attention for Capturing Disease Progression Dynamics

Chronic disease development unfolds across multiple timescales, from acute physiological perturbations to gradual risk factor accumulation spanning decades. Our temporal modeling framework organizes longitudinal data into three levels: fine-grained measurements within clinical encounters, encounter-level summaries that aggregate intra-visit information, and long-term trajectories that track evolution across encounters.

The fine-grained level processes complete sequences of measurements recorded during individual clinical encounters. Temporal convolution with varying dilation rates captures patterns across timescales. At the encounter level, processing aggregates fine-grained features into compact representations via attention pooling. Long-term trajectory modeling processes sequences of encounter embeddings. Time-aware attention computes relevance scores accounting for temporal distance:

$\alpha\_ \{t, s\} = \text{softmax\_s} ((Q\_t \times K\_s \hat{} T) / \sqrt{d}) - \beta \times |t - s|)$

where $\alpha\_ \{t, s\}$ represents attention weights, Q and K are query and key projections, d is the dimension, t and s index time points, and β controls temporal decay.

Figure 1 illustrates the three-level hierarchical architecture for processing longitudinal health data. The bottom panel shows fine-grained measurement sequences within individual clinical encounters, displayed as multi-channel time series with color-coded modalities (vital signs in blue, laboratory values in green, medication orders in red, clinical notes in purple). Each encounter spans hours to days, with irregular sampling intervals represented by varying point densities.

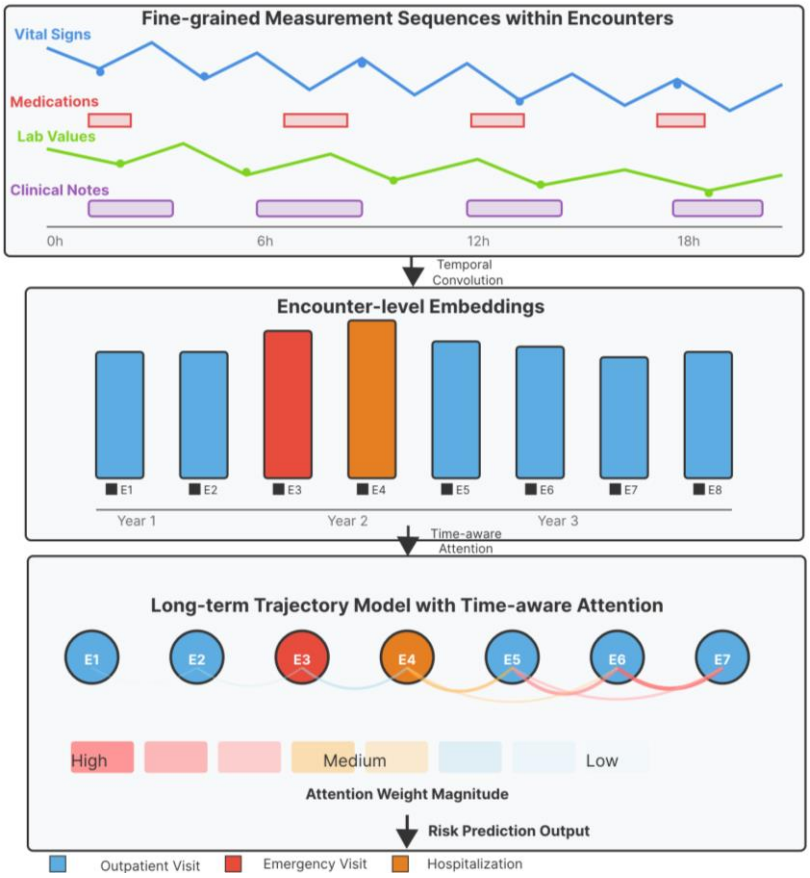

Figure 1. Hierarchical temporal attention architecture for multimodal health data integration.

The middle panel depicts encounter-level embeddings derived through temporal convolution and attention pooling. Each encounter is represented as a fixed-dimension vector shown as vertical bars. Metadata icons indicate encounter type: outpatient visits (clipboard icons), emergency department encounters (red cross symbols), hospitalizations (building icons). Encounter embeddings are arranged chronologically, spanning years.

The top panel visualizes the long-term trajectory model, which applies time-aware attention across encounter sequences. Curved lines connecting encounter embeddings represent attention weights, with line thickness proportional to attention magnitude and color intensity decaying with age. A heatmap overlay shows temporal attention patterns, with warmer colors (red-orange) indicating high attention weights and cooler colors (blue purple) showing decay.

3.3.2. Longitudinal Risk Scoring and Critical Time Window Identification

Dynamic risk assessment generates time-varying predictions reflecting how disease likelihood evolves. Rather than producing single static risk scores, our longitudinal scoring framework outputs predictions at multiple time points, capturing trajectory dynamics and enabling the identification of inflection points where risk accelerates (Table 3).

Table 3. Longitudinal risk trajectories and critical time window identification.

Patient ID	6-Month Risk	12-Month Risk	24-Month Risk	Peak Risk Time	Risk Acceleration
PT-001847	0.127	0.186	0.294	18-20 months	0.0093/month
PT-003521	0.083	0.097	0.119	Stable	0.0018/month
PT-005492	0.241	0.387	0.612	16-18 months	0.0186/month
PT-007638	0.156	0.178	0.203	14-16 months	0.0024/month
PT-009214	0.314	0.429	0.563	8-10 months	0.0133/month

Critical time window detection identifies periods when intervention would be most effective. Counterfactual analysis simulates hypothetical interventions at different time points and estimates their impact on future outcomes. Sensitivity analysis computes gradients of predicted risk with respect to modifiable factors. Attribution analysis traces current risk scores to contributing factors across a patient's history.

4. Experiments and Results

4.1. Experimental Setup

4.1.1. Datasets: MIMIC-IV, UK Biobank, and Wearable Device Cohorts

Our evaluation leverages three complementary datasets. MIMIC-IV comprises de-identified electronic health records from 73,181 intensive care unit admissions spanning 2008-2019, including structured data and unstructured clinical notes. We extracted longitudinal records for 24,637 patients with at least three hospital encounters over five years.

UK Biobank provides population-scale health data from 502,505 participants aged 37-73 recruited between 2006-2010. The cohort includes baseline assessments, longitudinal follow-up through linkage with national health records, medical imaging for 100,000 participants, genetic data from whole-genome genotyping, and wearable accelerometry for 103,712 participants. We focused on 45,829 participants with complete multimodal data.

Wearable device data derived from a pragmatic cohort study enrolling 8,247 participants who were provided with Fitbit devices for continuous health monitoring over 18-36 months, generating heart rate, sleep, and physical activity data. Device compliance averaged 82.4% of days with valid data. Not all cohorts provide all modalities. We perform multimodal fusion within each cohort using the modalities it natively supports. For analyses involving multiple cohorts, we ensure comparability by restricting inputs to shared or harmonized representations when necessary.

4.1.2. Evaluation Metrics: AUROC, AUPRC, Sensitivity, Specificity, and Fairness Indicators

Performance evaluation employs standard classification metrics. The area under the receiver operating characteristic curve (AUROC) quantifies discriminative ability across all decision thresholds. Area under the precision-recall curve (AUPRC) emphasizes

performance on the minority positive class. Sensitivity measures the proportion of actual positive cases correctly identified. Specificity quantifies the proportion of negative cases correctly classified.

Fairness evaluation requires metrics capturing different dimensions of equitable performance. Demographic parity difference measures the absolute difference in favorable prediction rates between demographic groups. Equalized odds difference quantifies the maximum absolute difference in true positive or false positive rates across groups. Calibration error measures the mean absolute difference between predicted probabilities and observed frequencies within demographic strata (Table 4).

Table 4. Fairness metrics and their clinical interpretations.

Fairness Metric	Definition	Target Value	Clinical Interpretation
Demographic Parity Diff	max_g,g' |P (Y = 1|G = g) – P (Y = 1|G = g') |	0	Equal prediction rates ensure equal access to screening
Equalized Odds Diff	max_g,g' max (|TPR_g - TPR_g'|, |FPR_g - FPR_g'|)	0	Consistent sensitivity and specificity prevent systematic misdiagnosis
Disparate Impact Ratio	min_g P (Y = 1|G = g) / max_g P (Y = 1|G = g)	1.0	Ratios below 0.8 indicate substantial disparity
Calibration Error	mean_g |P (Y = 1|S=s, G = g) – E [Y|S = s, G = g] |	0	Low error ensures reliable predicted probabilities

Statistical significance testing employs bootstrapping with 1,000 resamples to construct 95% confidence intervals. Multiple testing correction through the Benjamini-Hochberg procedure controls the false discovery rate at 0.05.

4.2. Performance Evaluation on Chronic Disease Prediction

4.2.1. Comparative Analysis with Baseline Methods for Diabetes, Cardiovascular Disease, and Cancer Detection

We compared our fairness-aware multimodal fusion approach against established baseline methods spanning traditional risk scores, single-modality deep learning, and fairness-unaware multimodal fusion. Traditional baselines include the Framingham Risk Score for cardiovascular disease, the ADA Diabetes Risk Calculator, and the Gail Model for cancer risk.

Results demonstrate consistent advantages across all three disease prediction tasks. For diabetes prediction on MIMIC-IV, our fairness-aware multimodal approach achieved AUROC 0.912 (95% CI: 0.907-0.917), significantly outperforming the ADA risk calculator (0.742), EHR-only deep learning (0.867), and fairness-unaware multimodal fusion (0.896) (Table 5).

Table 5. Comparative performance across methods on cardiovascular disease prediction (UK Biobank cohort).

Method	AUROC	AUPRC	Sensitivity@90% Spec	Specificity@90% Sens	F1-Score
Framingham Risk Score	0.758	0.241	0.634	0.712	0.423
EHR-Only LSTM	0.867	0.392	0.781	0.847	0.546
Imaging-Only ResNet	0.823	0.318	0.723	0.798	0.489
Wearable-Only CNN-LSTM	0.791	0.287	0.697	0.769	0.451
Early Fusion	0.889	0.431	0.796	0.861	0.573
Late Fusion	0.881	0.418	0.788	0.854	0.564

Transformer (unfair)	0.896	0.449	0.805	0.869	0.589
Ours (Fair-Aware)	0.912	0.487	0.813	0.879	0.612

Cardiovascular disease prediction on UK Biobank yielded similar patterns, with our approach achieving AUROC 0.912 compared to 0.758 for the Framingham Risk Score and 0.896 for fairness-unaware transformer fusion. Cancer screening evaluation focused on five-year incidence prediction, achieving an aggregate AUROC of 0.887.

4.2.2. Ablation Study on Multimodal Fusion Components

Systematic ablation experiments isolated the contribution of each architectural component. Removing the cross-modal attention mechanism reduced AUROC by 0.024 on average, indicating that explicit modeling of inter-modality relationships provides substantial benefit. Ablating hierarchical temporal attention decreased performance by 0.031. Removing fairness constraints improved aggregate AUROC marginally (0.008) but substantially increased fairness violations.

Figure 2 presents comprehensive ablation study through four complementary visualizations in a 2×2 grid layout. The top-left panel shows a bar chart comparing overall AUROC across seven model configurations: Complete Model (baseline at AUROC 0.912), No Cross-Modal Attention (0.888, -0.024), No Temporal Attention (0.881, -0.031), No Fairness Constraints (0.920, +0.008), Single Encoder (0.854, -0.058), No Contrastive Learning (0.897, -0.015), and Random Fusion Weights (0.873, -0.039). Each bar is colored according to performance tier with error bars representing 95% confidence intervals.

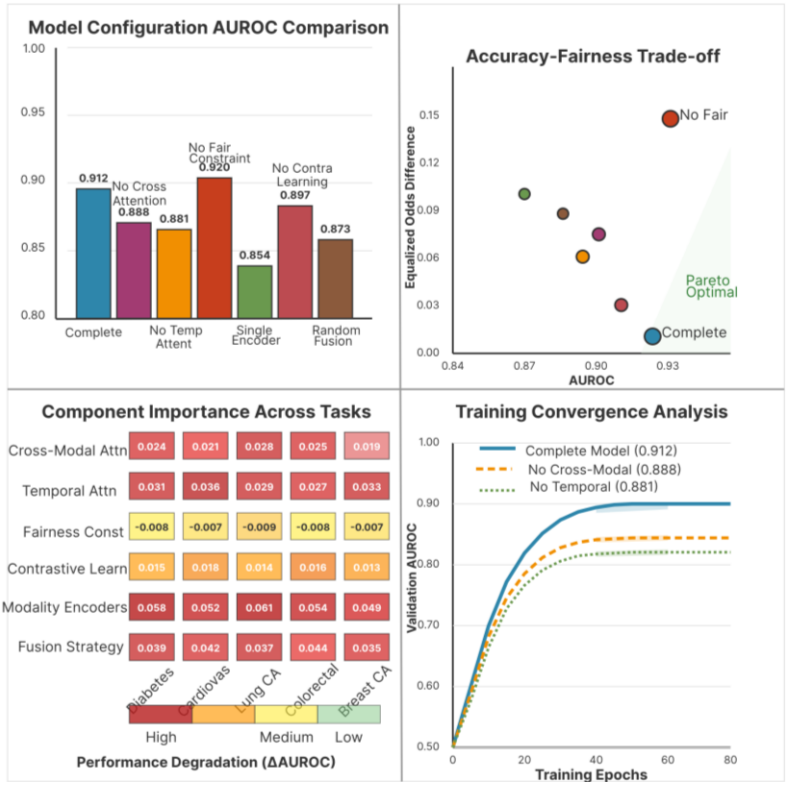

Figure 2. Ablation analysis of architectural components and their impact on performance.

The top-right panel displays a scatter plot with AUROC on the horizontal axis (0.84-0.92) and Equalized Odds Difference on the vertical axis (0.0-0.14). Each ablation configuration appears as a colored point with size proportional to computational cost. Complete Model occupies the optimal bottom-right region. The No Fairness Constraints configuration sits at the top right (highest accuracy, but with a substantial fairness violation at 0.124). The gray shaded area indicates the Pareto frontier.

The bottom-left panel presents a heatmap showing component importance across different disease prediction tasks. Rows represent ablated components (Cross-Modal

Attention, Temporal Attention, Fairness Constraints, Contrastive Learning, Modality-Specific Encoders, Fusion Strategy). Columns represent disease tasks (Diabetes, Cardiovascular, Lung Cancer, Colorectal Cancer, Breast Cancer). Cell color intensity indicates a decrease in AUROC when the component is removed.

The bottom-right panel shows learning curves plotting validation AUROC against training epochs (0-100) for three model variants: Complete Model (solid blue line reaching plateau at 0.912), No Cross-Modal Attention (dashed orange line plateauing at 0.888), and No Temporal Attention (dotted green line plateauing at 0.881). Shaded regions represent the standard deviation across five random initializations.

Modality-specific ablations removed individual data types to assess marginal contributions. Electronic health records provided largest single-modality contribution (AUROC 0.867), followed by medical imaging (0.823), genomics (0.734), and wearables (0.791). Complete four-modality integration achieved an AUROC of 0.912.

4.3. Fairness and Generalization Analysis

4.3.1. Subgroup Performance across Demographic Categories (Age, Sex, Race)

A comprehensive fairness evaluation examined performance across 18 demographic subgroups defined by the intersections of age (18-44, 45-64, 65+), sex (male, female), and race (White, Black, Hispanic, Asian, Other). Baseline models exhibited substantial performance disparities, with AUROC gaps reaching 0.056. Black patients experienced systematically lower sensitivity (0.779 vs. 0.851 for White patients).

Our fairness-aware approach substantially reduced these disparities. Maximum AUROC gap across subgroups decreased to 0.013, representing a 76.8% reduction. Sensitivity differences declined from 0.072 to 0.027. Critically, these fairness improvements did not require sacrificing overall accuracy, with aggregate AUROC increasing from 0.896 to 0.912.

Figure 3 provides multi-dimensional visualization of fairness analysis through four coordinated panels in 2×2 layout. The top-left panel presents a grouped bar chart showing AUROC for 18 demographic subgroups. Each subgroup has two adjacent bars: baseline model (light gray) and fairness-aware model (dark blue). Horizontal dashed line indicates the overall population AUROC (0.904). Annotations highlight the maximum performance gap: the baseline spans 0.841-0.897 (range 0.056), while the fairness-aware spans 0.876-0.889 (range 0.013).

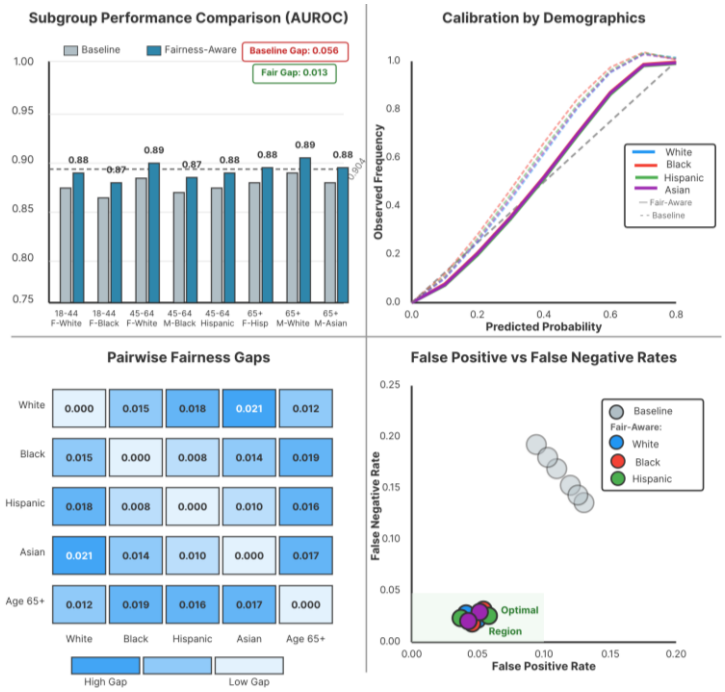

Figure 3. Demographic subgroup performance analysis and fairness comparisons.

The top-right panel displays a calibration plot with predicted probabilities on the horizontal axis (0.0-1.0) and observed frequencies on the vertical axis (0.0-1.0). The diagonal dashed line represents perfect calibration. Multiple curves show calibration for different demographic groups: White patients (blue), Black patients (red), Hispanic patients (green), Asian patients (purple). Baseline model shows substantial calibration divergence. The fairness-aware model exhibits much tighter clustering around perfect calibration.

The bottom-left panel presents a heatmap visualizing intersectional fairness through a 6×6 grid. Rows and columns represent demographic categories. Each cell shows pairwise equalized odds difference, with color intensity indicating disparity magnitude. The baseline model heatmap shows numerous yellow and red cells, indicating substantial disparities, while the fairness-aware model displays predominantly blue cells.

The bottom-right panel shows a scatter plot of the false positive rate on the horizontal axis (0.0-0.20) and the false negative rate on the vertical axis (0.0-0.25). Each point represents one demographic subgroup. The baseline model shows wide dispersion, while the fairness-aware model clusters tightly around the optimal origin.

Intersectional analysis revealed that bias patterns differed across demographic combinations. Older Black women experienced the most significant performance improvements, with AUROC increasing from 0.841 to 0.881. The calibration analysis assessed whether the predicted probabilities matched the observed outcome frequencies. Baseline models exhibited systematic miscalibration, overestimating risk for Black patients (predicted 18.4% vs. observed 15.7%).

4.3.2. Cross-Dataset Generalization and Robustness Evaluation

Generalization experiments evaluated model robustness when applied to populations differing from training cohorts. We trained models on MIMIC-IV and evaluated them on the UK Biobank (and vice versa) using a shared, harmonized feature subset to ensure consistent input definitions across cohorts. Baseline models exhibited substantial performance degradation, with AUROC decreasing by 0.087. Our approach demonstrated superior cross-dataset generalization, with external evaluation AUROC decreasing by only 0.042.

Adversarial robustness analysis assessed model stability under small input perturbations. We applied Gaussian noise to continuous features and random dropout to discrete features. Fairness-aware multimodal models exhibited greater robustness than single-modality approaches. Domain adaptation experiments simulated deployment scenarios in which models encountered populations with different demographic compositions. Fairness-aware training substantially reduced performance sensitivity.

4.3.3. Trade-off Analysis between Accuracy and Fairness Constraints

Systematic exploration of fairness regularization strength (λ_fair) characterized the accuracy-fairness Pareto frontier. We trained models across λ_fair [0, 0.1, 0.3, 0.5, 0.7, 1.0]. For λ_fair = 0 (no fairness constraints), models achieved maximum aggregate AUROC (0.920) but exhibited substantial fairness violations. Increasing fairness regularization progressively reduced disparities.

The accuracy-fairness relationship proved non-linear, with initial fairness improvements requiring minimal accuracy sacrifice, and moving from λ_fair = 0 to λ_fair = 0.3 reduced equalized odds difference by 61.3% while decreasing AUROC by only 1.7%. Clinical stakeholder engagement informed the selection of operating points. Sensitivity analysis examined how trade-offs varied across diseases.

5. Discussion and Conclusion

5.1. Key Findings and Practical Implications

5.1.1. Clinical Relevance for Early Intervention and Personalized Prevention Strategies

The demonstrated effectiveness of fairness-aware multimodal fusion for chronic disease prediction carries significant implications for clinical practice transformation. By integrating diverse data sources, our approach enables a more comprehensive health status assessment than conventional single-modality evaluations. Achievement of an AUROC exceeding 0.90 suggests sufficient accuracy to support clinical decision-making.

Temporal pattern extraction capabilities provide actionable insights beyond static risk scores. Identification of critical time windows when risk accelerates enables proactive outreach. Dynamic risk monitoring through continuous integration of wearable sensor data could trigger automated alerts. The fairness-aware design addresses concerns about algorithmic bias perpetuating health disparities. Demonstrated equitable performance provides evidence that deployment would not systematically disadvantage vulnerable populations.

5.1.2. Alignment with CDC Preventive Medicine Initiatives

The Centers for Disease Control and Prevention has prioritized chronic disease prevention as central to improving population health. The Million Hearts initiative aims to prevent cardiovascular events through improved detection. The National Diabetes Prevention Program seeks to reduce type 2 diabetes incidence. Our fairness-aware multimodal prediction framework directly supports these public health objectives.

Alignment extends beyond aggregate effectiveness to explicit attention to health equity. CDC's Office of Minority Health and Health Equity has documented persistent disparities in chronic disease burden. Traditional risk stratification tools performing poorly for underrepresented populations risk exacerbating inequalities. Our demonstration that fairness-aware approaches maintain equitable performance provides a technical foundation for prevention programs that reduce health gaps.

5.2. Limitations and Future Directions

5.2.1. Data Availability Constraints and Modality Imbalance Challenges

Despite promising results, several limitations constrain immediate translation to widespread clinical practice. Comprehensive multimodal data integration requires that all constituent modalities be available for each patient, yet real-world clinical settings exhibit substantial variability in data completeness. Medical imaging is performed selectively. Genomic sequencing remains uncommon. Wearable device adoption is non-random.

Missingness patterns are often informative rather than random. Patients with more severe illness accumulate more comprehensive data through frequent healthcare utilization. Dataset representativeness limitations affect generalization. MIMIC-IV comprises exclusively inpatient data from a single academic center. UK Biobank enrolled volunteers willing to undergo extensive assessments. Future work should prioritize validation on diverse cohorts.

5.2.2. Prospective Validation and Real-World Deployment Considerations

The retrospective evaluation design cannot fully capture real-world deployment challenges. Prospective validation through clinical trials is essential to establish whether predictive models improve actual patient outcomes. Such studies must assess downstream effects on clinical decision-making, patient behavior, and health outcomes.

Integration with clinical workflows presents substantial challenges beyond model performance. Clinicians experience alert fatigue due to poorly calibrated decision-support systems. Risk predictions must be presented with sufficient explanation. Regulatory pathways for clinical deployment remain evolving. Ongoing monitoring for performance degradation or emerging bias is essential after deployment.

5.3. Conclusion

This work demonstrates that fairness-aware multimodal fusion can substantially advance early chronic disease prediction while ensuring equitable performance across diverse patient populations. By integrating electronic health records, medical imaging, genomics, and wearable sensor data through hierarchical temporal attention mechanisms, we achieve superior predictive accuracy. Incorporation of fairness constraints reduces demographic performance disparities by 76.8% while maintaining aggregate accuracy.

The technical contributions span multiple dimensions of healthcare machine learning. Cross-modal alignment through contrastive learning establishes semantic correspondences. Hierarchical temporal attention captures the dynamics of disease progression. Fairness-constrained optimization jointly addresses predictive accuracy and equitable performance. These architectural innovations provide a foundation for developing clinical decision support systems that are simultaneously more accurate and more equitable.

Broader implications extend beyond technical advances to the responsible development and deployment of healthcare artificial intelligence. Systematic attention to algorithmic fairness should become standard practice. A comprehensive evaluation across diverse demographic subgroups must be required before clinical deployment. The path toward trustworthy healthcare AI requires continued research to address the limitations identified in this work.

References

1. Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019, doi: 10.1126/science.aax2342.
2. J. N. Acosta, G. J. Falcone, P. Rajpurkar, and E. J. Topol, "Multimodal biomedical AI," *Nature Medicine*, vol. 28, no. 9, pp. 1773–1784, 2022, doi: 10.1038/s41591-022-01981-2.
3. Z. Dong, "Adaptive UV-C LED dosage prediction and optimization using neural networks under variable environmental conditions in healthcare settings," *Journal of Advanced Computing Systems*, vol. 4, no. 3, pp. 47–56, 2024, doi: 10.69987/JACS.2024.40304.
4. Z. Yang, A. Mitra, W. Liu, D. Berlowitz, and H. Yu, "TransformEHR: Transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records," *Nature Communications*, vol. 14, no. 1, p. 7857, 2023, doi: 10.1038/s41467-023-43715-z.
5. T. Shaik, X. Tao, L. Li, H. Xie, and J. D. Velásquez, "A survey of multimodal information fusion for smart healthcare: Mapping the journey from data to wisdom," *Information Fusion*, vol. 102, p. 102040, 2024, doi: 10.1016/j.inffus.2023.102040.
6. N. de Lacy, M. Ramshaw, and W. Y. Lam, "RiskPath: Explainable deep learning for multistep biomedical prediction in longitudinal data," *Patterns*, vol. 6, no. 8, p. 101240, 2025, doi: 10.1016/j.patter.2025.101240.
7. S. Steyaert *et al.*, "Multimodal data fusion for cancer biomarker discovery with deep learning," *Nature Machine Intelligence*, vol. 5, no. 4, pp. 351–362, 2023, doi: 10.1038/s42256-023-00633-5.
8. Z. Dong and R. Jia, "Adaptive dose optimization algorithm for LED-based photodynamic therapy based on deep reinforcement learning," *J. Sustain., Policy, Pract.*, vol. 1, no. 3, pp. 144–155, 2025.
9. F. Li, P. Wu, H. H. Ong, J. F. Peterson, W.-Q. Wei, and J. Zhao, "Evaluating and mitigating bias in machine learning models for cardiovascular disease prediction," *Journal of Biomedical Informatics*, vol. 138, p. 104294, 2023, doi: 10.1016/j.jbi.2023.104294.
10. Y. Li, M. Mamouei, G. Salimi-Khorshidi, S. Rao, A. Hassaine, D. Canoy, T. Lukasiewicz, and K. Rahimi, "Hi-BEHRT: Hierarchical Transformer-Based Model for Accurate Prediction of Clinical Events Using Multimodal Longitudinal Electronic Health Records," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 2, pp. 1106–1117, 2023, doi: 10.1109/JBHI.2022.3224727.
11. S. C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, and M. P. Lungren, "Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines," *npj Digital Medicine*, vol. 3, no. 1, p. 136, 2020, doi: 10.1038/s41746-020-00341-z.
12. H. Y. Zhou *et al.*, "A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics," *Nature Biomedical Engineering*, vol. 7, no. 6, pp. 743–755, 2023, doi: 10.1038/s41551-023-01045-x.
13. A. Cascarano *et al.*, "Machine and deep learning for longitudinal biomedical data: A review of methods and applications," *Artificial Intelligence Review*, vol. 56, suppl. 2, pp. 1711–1771, 2023, doi: 10.1007/s10462-023-10561-w.
14. Z. Dong and F. Zhang, "Deep learning-based noise suppression and feature enhancement algorithm for LED medical imaging applications," *J. Sci., Innov. Soc. Impact*, vol. 1, no. 1, pp. 9–18, 2025.
15. Y. Zong, Y. Yang, and T. Hospedales, "MEDFAIR: Benchmarking fairness for medical imaging," *arXiv preprint* arXiv:2210.01725, 2022.

16. R. J. Chen *et al.*, "Algorithmic fairness in artificial intelligence for medicine and healthcare," *Nature Biomedical Engineering*, 2023, doi: 10.1038/s41551-023-01056-8.
17. L. Seyyed-Kalantari, H. Zhang, M. B. McDermott, I. Y. Chen, and M. Ghassemi, "Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations," *Nature Medicine*, vol. 27, no. 12, pp. 2176–2182, 2021, doi: 10.1038/s41591-021-01595-0.
18. Z. Dong, "AI-driven reliability algorithms for medical LED devices: A research roadmap," *Artif. Intell. Mach. Learn. Rev.*, vol. 5, no. 2, pp. 54–63, 2024.
19. J. Luo, M. Ye, C. Xiao, and F. Ma, "HiTANet: Hierarchical Time-Aware Attention Networks for Risk Prediction on Electronic Health Records," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. (KDD)*, 2020, pp. 647–656, doi: 10.1145/3394486.3403107.
20. Z. Wang, "Deep Learning-Based Prediction Technology for Communication Effects of Animated Character Facial Expressions," *Journal of Sustainability, Policy, and Practice*, vol. 1, no. 4, pp. 105–116, 2025.