Journal of Current Research and Studies

Volume 03 Issue 04 July-August 2026
https://journalcurrentresearch.com/
ISSN:3048-9911

OPEN ACCESS

***Correspondence**
Muhammed Kuliya

Article Received
13/06/2026
Accepted
24/06/2026
Published
10/07/2026

Works Cited

Muhammed Kuliya, Zaharaddeen Salele Iro., (2026). Adaptive Multi-head Attention and Residual Gated Linear Units: An Ablation Analysis of Transformer Enhancements for Multimodal Clinical Tabular Data. *Journal of Current Research and Studies*, 3(4), 25-40.

***COPYRIGHT**

© 2026 Muhammed Kuliya. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms

Adaptive Multi-head Attention and Residual Gated Linear Units: An Ablation Analysis of Transformer Enhancements for Multimodal Clinical Tabular Data

Muhammed Kuliya[1] & Zaharaddeen Salele Iro[2]
[1,2]Department of Information Technology Science, Federal University Dutse, Dutse, Jigawa State, Nigeria

Abstract

Background: Transformer architectures for tabular data have shown promise, but their optimal design for multimodal cardiovascular risk prediction remains unclear. This proof-of-concept study systematically ablates two architectural innovations using a synthetically aligned multimodal dataset derived from NHANES, Framingham, and Kaggle sources.
Objective: To systematically ablate two architectural innovations, Adaptive Multi-Head Attention (with learnable scaling parameters α, β) and Residual Gated Linear Units (ResidualGLU) within a TabTransformer framework for multitask prediction of coronary heart disease (CHD) and stroke.
Methods: Controlled ablation on fused multimodal dataset (N=10,410, 21 features from physiological, environmental, demographic, behavioral modalities). Four configurations: (1) Baseline TabTransformer, (2) +Adaptive Attention only, (3) +ResidualGLU only, (4) Full EnTabTransformer (both). Evaluation: 5-fold cross-validation with accuracy, F1-score, AUC-ROC. Statistical significance via paired t-tests with Bonferroni correction. Mechanistic analysis: gradient flow tracking (30 runs) and attention entropy.
Results: Adaptive Attention alone improved CHD AUC by +1.31pp ($p<0.001$) and stroke AUC by +1.75pp ($p<0.001$). ResidualGLU alone yielded +0.76pp (CHD, $p=0.012$) and +1.02pp (stroke, $p=0.008$). Full EnTabTransformer achieved super-additive gains: +2.71pp (CHD) and +3.77pp (stroke) exceeding sum of individual improvements by 31% (CHD) and 36% (stroke). Gradient analysis showed ResidualGLU reduced vanishing gradient probability by 34% ($p=0.003$). Adaptive Attention increased attention entropy by 0.42 bits ($p<0.001$), indicating broader feature coverage. Computational overhead: +18% parameters, +10% inference time. All results are based on a proof-of-concept dataset that includes synthetic ECG and simulated patient alignment; prospective validation on authentic multimodal clinical data is required before deployment.
Conclusion: Adaptive Attention and ResidualGLU address orthogonal aspects of feature representation, cross-feature interaction weighting vs. non-linear transformation capacity. Their combination yields synergetic gains. For resource-constrained deployment, Adaptive Attention alone is recommended; for maximum accuracy, full EnTabTransformer is justified.

Keywords: Transformer ablation, multimodal clinical data, cardiovascular risk prediction, adaptive attention, gated linear units, tabular deep learning

1.0 Introduction

The rapid proliferation of Internet-of-Things (IoT) sensors, electronic health records (EHRs), and wearable devices has ushered in a new era of data-driven cardiovascular medicine (Moshawrab et al., 2023; Teoh et al., 2024). Smart healthcare systems now routinely collect multimodal patient data covering physiological signals, environmental, demographic characteristics, and behavioral patterns creating unprecedented opportunities for predictive modeling. However, these rich data streams are heterogeneous, high-dimensional, and characterized by complex, non-linear interactions that conventional machine learning models struggle to capture effectively (Zhao et al., 2024).

Transformer architectures, originally developed for natural language processing (Vaswani et al., 2017), have recently demonstrated remarkable adaptability for time-series, multimodal, and tabular data problems (Noor et al., 2025). In the healthcare domain, transformer-based models have shown promise for tasks ranging from disease diagnosis to risk stratification (Kline et al., 2022; Madan et al., 2024). The self-attention mechanism that supports these architectures enables the modeling of long-range dependencies and dynamic feature weighting capabilities that are particularly valuable for clinical data where feature relevance is often context-dependent.

The TabTransformer (Huang et al., 2020), specifically designed for tabular data, processes categorical features through embedding layers and applies transformer encoder blocks to capture complex feature interactions. When combined with numerical features, this architecture has shown strong performance across finance, retail, and increasingly, healthcare applications (Alam et al., 2023). For cardiovascular disease prediction tasks, transformer-based approaches have achieved accuracies comparable to or exceeding traditional ensemble methods (Dubey et al., 2025; Talaat & Aly, 2025).

Despite these advances, the optimal design of transformer architectures for multimodal clinical tabular data remains poorly understood. Two critical questions persist: First, how do specific architectural innovations, particularly enhancements to the attention mechanism and the feed-forward network individually and jointly contribute to predictive performance? Additionally, what are the mechanistic bases for any observed improvements, and how should practitioners trade off accuracy against computational efficiency in resource-constrained settings?

This study systematically ablates two architectural innovations within a TabTransformer framework for multitask cardiovascular risk prediction (CHD and stroke):

i. Adaptive Multi-Head Attention with learnable scaling parameters (α, β) that enable context-dependent sharpening or flattening of attention distributions; a capability that may be particularly valuable when clinical feature relevance varies across patient subgroups or disease contexts.
ii. ResidualGLU that replace standard feed-forward networks with gated mechanisms, building on prior work demonstrating the effectiveness of gated linear units for controlling information flow in neural networks (Shazeer, 2020; Y. Zheng et al., 2022).

Our contributions are fourfold:

i. First systematic ablation of Adaptive Attention + ResidualGLU for clinical tabular data, providing empirical evidence for the individual and combined contributions of each component.
ii. Demonstration of super-additive gains proving architectural synergy, with combined effects exceeding the sum of individual improvements by 31–36%.
iii. Mechanistic analysis via gradient flow tracking and attention entropy, revealing distinct roles: ResidualGLU improves gradient propagation, while Adaptive Attention increases feature coverage.
iv. Practical design guidelines enabling informed choices between resource-efficient and maximum-accuracy configurations for clinical deployment.

We emphasize that this study is a methodological proof-of-concept. The dataset used involves simulated patient alignment across separate cohorts and synthetically generated ECG signals. Our goal is to evaluate architectural synergies under controlled conditions, not to claim clinical readiness. All medical language should be understood within this proof-of-concept context.

Beyond architectural evaluation, this study provides a deployable real-time implementation via Streamlit (average inference latency 297ms) and quantifies modality-specific efficiency ratios addressing the translational gap where most prior work stops at performance metrics without delivering a usable clinical tool.

The remainder of this paper is organized as follows. Section 2 reviews related work on transformer architectures for clinical data, attention mechanisms, and gated linear units. Section 3 describes our methodology, including dataset characteristics, baseline architecture, proposed enhancements, and evaluation protocols. Section 4 presents experimental results across four configurations, including statistical significance testing, gradient analysis, and attention entropy evaluation. Section 5 interprets these findings, discusses mechanistic insights, and provides practical guidelines. Section 6 concludes with implications for future research.

2.0 Background and Related Work

2.1 Transformer Architectures for Tabular Healthcare Data

The transformer architecture, introduced by Vaswani et al. (2017), revolutionized sequence modeling through its self-attention mechanism, which computes contextualized representations by weighting all elements in a sequence according to their relevance. Unlike recurrent neural networks, transformers process all elements in parallel, enabling efficient training and capturing long-range dependencies.

For tabular data, where features lack an inherent sequential order, several adaptations have emerged. The TabTransformer (Huang et al., 2020) embeds categorical features into a latent space and applies transformer encoder blocks to model interactions among these embeddings, then concatenates the output with numerical features for final prediction. This approach has proven effective across diverse domains. For instance, Alam et al. (2023) applied TabTransformer to predict hospital length of stay for COVID-19 patients, demonstrating superior performance compared to traditional machine learning methods. In cardiovascular applications, the CardioTabNet model (Shaheenur Islam Sumon et al., 2025) leverages TabTransformer for heart disease prediction, highlighting the importance of clinical categorical variables such as electrocardiogram findings and exercise-induced angina.

The FT-Transformer (Feature Tokenizer Transformer) (Gorishniy et al., 2023) offers an alternative paradigm, treating both numerical and categorical features as tokens that are projected into a common embedding space and processed through standard transformer encoder blocks. This unified treatment of all features has shown strong generalization across tabular machine learning benchmarks.

SAINT (Self-Attention and Intersample Attention Transformer) (Somepalli et al., 2021) introduces a dual-attention mechanism that captures dependencies both among features (intra-sample attention) and across different data instances (inter-sample attention), further enhancing predictive performance for classification and regression tasks.

Despite these advances, the application of transformer architectures to multimodal clinical data, where physiological signals, environmental factors, and demographic information must be integrated remains emerging. Recent work by Noor et al. (2025) demonstrated a transformer-based approach for cardiovascular disease classification from electrocardiography recordings, achieving impressive performance metrics but focusing on unimodal data. Similarly, the Cardiac Sensing Foundation Model (CSFM) (Gu et al., 2026) leverages transformer architectures pretrained on multimodal data from approximately 1.7 million individuals, highlighting the growing recognition of transformers' potential for unified cardiac health assessment.

A systematic review of transformer-based models for CVD prediction from EHRs (Chikumo & Ndlovu, 2026) identified key opportunities, including federated learning and hybrid architectures that integrate transformers with traditional machine learning methods. However, this review also noted that most existing studies focus on performance reporting without systematic architectural ablation, leaving design choices largely heuristic.

2.2 Attention Mechanisms in Clinical Models

The multi-head self-attention mechanism (Vaswani et al., 2017) computes attention scores as:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \ldots \quad (1)$$

where Q, K and V are query, key, and value matrices derived from input embeddings. This mechanism allows the model to dynamically weight the importance of different positions in the input sequence.

In healthcare applications, attention mechanisms have been adapted to address domain-specific challenges. (Y. Zhang & Li, 2025) integrates temporal embeddings and hierarchical attention to model longitudinal patient data, showing substantial improvements on mortality prediction, readmission prediction, and comorbidity onset tasks. (Koo et al., 2025) employs multi-Head attention to capture complex dependencies among genes for patient survival prediction, demonstrating the value of multitask learning frameworks.

Recent work has explored adaptive attention mechanisms for medical data. The Adaptive Noise-Augmented Attention (ANAA) mechanism (Amirahmadi et al., 2025) broadens attention distribution across tokens while refining focus on informative events, operating entirely during fine-tuning without requiring expensive architectural modifications. (Zhang et al., 2025) introduces a dispatcher dual-attention mechanism with learnable dispatcher tokens serving as global communicators for biomedical irregular time series classification. (Ye et al., 2026) incorporates sparse

token-based dual-attention for medical time series, enabling dynamic feature refinement by focusing on informative tokens while reducing redundancy.
However, none of these approaches has systematically investigated the impact of learnable scaling parameters within the attention mechanism for multimodal clinical tabular data. Our work addresses this gap by introducing Adaptive Multi-Head Attention with learnable parameters α (scale) and β (shift), enabling context-dependent attention distribution sharpening.

2.3 Gated Linear Units in Deep Learning

Gated Linear Units (GLUs), introduced by Dauphin et al. (2017), are neural network layers defined as:

$$GLU(x) = (xW_1 + b_1) \otimes \sigma(xW_2 + b_2)\ldots \quad (2)$$

where $\otimes$ denotes element-wise multiplication and σ is the sigmoid activation function. The gating mechanism allows the network to control information flow adaptively, suppressing irrelevant features while amplifying important ones.
(Shazeer, 2020) systematically evaluated GLU variants within transformer architectures, demonstrating that replacing standard feed-forward networks with GLU-based alternatives consistently improves performance across multiple tasks. The study identified that GLU variants with sigmoid activation achieve strong results while maintaining computational efficiency.
In healthcare applications, GLUs have shown particular promise for handling noisy physiological signals. (Le et al., 2025) investigated Gated Residual Networks within transformers for photoplethysmogram artifact detection in pediatric intensive care units, finding that GLU with sigmoid activation achieved 0.98 accuracy, 0.91 precision, 0.96 recall, and 0.94 F1-score. Their mutual information neural estimation analysis supported the hypothesis that gating mechanisms enhance the mutual information between hidden representations and outputs.
The combination of residual connections with GLUs forming ResidualGLU, has been explored in several contexts. Gated Residual Networks (GRN) (Lim et al., 2021) incorporate skip connections and gating layers to facilitate efficient information flow, applying GLU and adding the original input to the output before layer normalization. This residual structure facilitates gradient flow and prevents feature drift during training, enhancing model robustness (Z. Zheng et al., 2026).
Our work extends these investigations by systematically ablating ResidualGLU within a TabTransformer architecture for multitask cardiovascular risk prediction, quantifying its marginal contribution both independently and in combination with adaptive attention mechanisms.

2.4 Research Gap

A critical review of the literature reveals several gaps that motivate this study:

- At first, while transformer architectures have been applied to clinical tabular data, no prior study has systematically ablated the contributions of attention mechanism enhancements and feed-forward network modifications for multimodal cardiovascular risk prediction.
- Additionally, it remains unclear whether the gains from different architectural innovations are additive or synergistic. Understanding this distinction is crucial for model design: if gains are simply additive, practitioners can combine components without concern; if synergistic, the combined architecture may offer disproportionate benefits.
- Furthermore, mechanistic understanding on why specific innovations improve performance is largely absent from the literature. Without such understanding, architectural choices remain heuristic rather than principled.
- Also, practical design guidelines that help practitioners trade off accuracy against computational efficiency are needed, particularly for resource-constrained clinical deployment scenarios.

This study addresses these gaps through systematic ablation, statistical rigor, mechanistic analysis, and actionable recommendations.

3.0 Methods

3.1 Dataset Description

The dataset used in this study was constructed through fusion of four complementary sources covering:

i. Physiological data (synthetic ECG, N=10,410) comprising 8 key metrics including ECG_Rate, ECG_PR_Interval, ECG_QRS_Duration, ECG_QT_Interval, and ECG_Arrhythmia labels.
ii. Environmental data (N=1,000 records) with five features capturing air quality index, PM2.5 concentrations, noise pollution, green space access, and walkability score.
iii. Lifestyle and behavioral data from NHANES 2017-2018, focusing on tobacco use history.
iv. Disease outcome labels from Framingham Heart Study dataset (10-year CHD risk) and Kaggle Stroke Prediction dataset.

After preprocessing, including missing value imputation (mean for numerical, mode for categorical), z-score standardization, categorical encoding (one-hot for nominal, label for ordinal), and SMOTE-based class balancing, the final fused dataset comprised 10,410 samples with 21 optimally selected features. Feature selection was performed using a hybrid strategy combining ANOVA F-test and Random Forest importance, applied exclusively to the training set to prevent data leakage.

The dataset was split into training (80%), validation (15%), and test (20%) sets using stratified sampling based on both target variables (CHD and stroke). The class distribution post-SMOTE was balanced: 50.2% positive for CHD and 49.8% positive for stroke.

3.1.1 Data Transparency Statement

ECG signals were synthetically generated using the ecg_simulator Python library with parameters set to match clinical population norms. While this enables controlled methodological development and privacy-preserving initial validation, the reported performance metrics (particularly the high CHD AUC of 0.9843) should be interpreted as proof-of-concept upper-bound estimates. The fusion of datasets from different sources (NHANES, Framingham, Kaggle, synthetic ECG) required simulated patient alignment based on demographic characteristics, as no common unique identifier existed across these heterogeneous cohorts. Prospective validation on authentic, linked multimodal clinical data is required before clinical deployment.

3.2 Baseline TabTransformer Architecture

Our baseline TabTransformer follows the architecture described by (Huang et al., 2020). Key hyperparameters were (see table 1):

Parameter	Value
Embedding dimension (d_model)	128
Number of attention heads (h)	8
Number of encoder blocks (L)	4
Feed-forward dimension	512
Dropout rate	0.1
Positional encoding	Learnable

Table 1: Key parameters

Input processing: Categorical features are mapped to learnable embeddings of dimension 128. Numerical features are normalized (z-score) and projected through a linear layer to the same dimension. The sequence of feature embeddings (length = number of features, 21) is then processed through L transformer encoder blocks.

Transformer encoder block: Each block contains:

- Multi-head self-attention with residual connection and layer normalization
- Position-wise feed-forward network (two linear layers with GELU activation) with residual connection and layer normalization

Classification head: Global average pooling over the feature dimension, followed by a linear layer with sigmoid activation for binary classification. Separate heads are used for CHD and stroke prediction (multitask learning with shared lower layers).

3.3 Proposed Architectural Enhancements

3.3.1 Adaptive Multi-Head Attention

Standard multi-head attention uses a fixed scaling factor of $\frac{1}{\sqrt{d_k}}$ (see equation 1). We introduce learnable scaling parameters α (scale) and β (shift):

$$Attention_{adaptive}(Q,K,V) = softmax(\frac{\alpha.QK^T}{\sqrt{d_k}} + \beta)V \dots (3)$$

where α and β are learnable scalars initialized to α=1.0, β=0.0. These parameters are jointly optimized with the rest of the model during training.

The fixed scaling factor assumes that the optimal attention sharpness is constant across all contexts. However, in clinical data, the relevance of features varies by patient subgroup and disease context. For example, when predicting CHD in elderly patients, age and blood pressure may dominate; in younger patients, lifestyle factors may be more informative. Learnable α and β allow the model to dynamically adjust attention sharpness based on the input context where high α produces sharper attention (focusing on few features), low α produces flatter attention (distributing more evenly).

3.3.2 ResidualGLU

The standard position-wise feed-forward network in transformers is:

$$FFN(x) = GELU(xW_1 + b_1)W_2 + b_2 \cdots \quad (4)$$

We replace this with a ResidualGLU:

$$ResidualGLU(Z) = Z + Dropout(Linear_1(Z) \otimes \sigma)(Linear_2(Z))) \cdots \quad (5)$$

where:

- $Linear_1$ and $Linear_2$ are linear transformations (typically with output dimension matching the hidden dimension).
- $\otimes$ denotes element-wise multiplication.
- σ is the sigmoid activation function.
- The residual connection adds the original input Z to the gated output.

The gating mechanism $(\sigma(Linear_2(Z)))$ produces values between 0 and 1, acting as a soft gate that controls how much of the transformed information $(Linear_1(Z)$ is passed through. This allows the network to selectively suppress irrelevant or noisy information, a valuable property for handling physiological signals that may contain artifacts, missing values, or measurement errors. The residual connection preserves gradient flow, mitigating the vanishing gradient problem as network depth increases (He et al., 2016).

3.4 Experimental Configurations

We evaluate four configurations using controlled ablation (see table 2):

Configuration	Adaptive Attention	ResidualGLU	Abbreviation
1	✗	✗	Baseline
2	✓	✗	+Attn
3	✗	✓	+GLU
4	✓	✓	Full

Table 2: Experimental configurations for ablation study

All configurations share identical training protocols, hyperparameters, and data splits, ensuring that observed differences are attributable solely to the architectural variations.

3.5 Training Protocol

- Optimization: Adam optimizer with learning rate = 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay = 1e - 5.
- Batch size: 64
- Maximum epochs: 50 with early stopping (patience = 10 epochs based on validation loss)
- Loss function: Binary cross-entropy, applied separately to CHD and stroke outputs:

$$L_{BCE} = -\frac{1}{n}\sum_{i=1}^{n}[y_1 \log(\hat{y_i}) + (1 - y_i)\log(1 - \hat{y_i})] \cdots \quad (6)$$

- Gradient clipping: max_norm = 1.0
- Learning rate scheduling: Cosine annealing with warm restarts (T_o = 10 epochs)

3.6 Evaluation Metrics

- Primary metric: AUC-ROC (Area Under the Receiver Operating Characteristic Curve), chosen for its threshold-independence and sensitivity to the model's ability to rank patients by risk.
- Secondary metrics: Accuracy, F1-score (harmonic mean of precision and recall), Precision, Recall.
- Cross-validation: 5-fold stratified cross-validation, with performance reported as mean ± standard deviation across folds.
- Statistical significance: Paired t-tests comparing each configuration to the baseline, with Bonferroni correction for multiple comparisons ($\alpha = \frac{0.05}{6} = 0.0083$).

3.7 Mechanistic Analysis

3.7.1 Gradient Flow Analysis

To quantify the impact of ResidualGLU on gradient propagation, we tracked gradient norms across the four encoder layers during training:

- For each of 30 independent training runs (different random seeds), we recorded the gradient norm $||\nabla W||_2$ for the first and last transformer blocks at epoch 20 (post-convergence).
- Computed gradient decay ratio:

$$Decay = ||\nabla W_{layer4}\ ||/||\ \nabla W_{layer1}||$$

- Calculated vanishing gradient probability:

$$Pr(||\ \nabla W_{layer4}\ || < 0.01 \times ||\ \nabla W_{layer1}\ ||)$$

3.7.2 Attention Entropy Analysis

To understand how Adaptive Attention changes information flow:

- Extracted attention weight matrices from the first encoder block for 1,000 validation samples
- Computed **attention entropy** per sample:

$$H = -\sum_i p_i \log_2 p_i$$

where p_i are normalized attention weights across the 21 features.

- Normalized by $\log_2(21)$ to obtain values in [0,1]
- Compared distributions between standard and adaptive attention using two-sample t-test

3.7.3 Computational Overhead

- Counted trainable parameters for each configuration
- Measured inference latency on CPU (Intel (R) Core (TM) i5 – 7200u CPU @ 2.50GHz 2.71GHz) over 1,000 sequential requests, reporting mean ± standard deviation

4.0 Results

4.1 Baseline TabTransformer Performance

The baseline TabTransformer (without architectural enhancements) achieved solid performance across both prediction tasks, serving as a reference point for ablation comparisons (see table 3).

Metric	CHD	Stroke
Accuracy	0.9213 ± 0.0031	0.8712 ± 0.0042
Precision	0.9235 ± 0.0034	0.8621 ± 0.0048
Recall	0.9219 ± 0.0032	0.8685 ± 0.0043
F1-Score	0.9227 ± 0.0028	0.8653 ± 0.0039
AUC-ROC	0.9572 ± 0.0025	0.9112 ± 0.0038

TABLE 3: Baseline TabTransformer Performance (Mean ± SD, 5-Fold CV)

These results confirm that the baseline TabTransformer is a strong performer for cardiovascular risk prediction, with particularly strong discrimination for CHD (AUC 0.9572). The lower stroke AUC (0.9112) reflects the etiological complexity of stroke, which involves multiple mechanisms (thrombotic, embolic, hemorrhagic) that may be more difficult to capture from tabular data alone.

4.2 Ablation: Adaptive Multi-Head Attention Only

Adding Adaptive Multi-Head Attention to the baseline produced substantial, statistically significant improvements across all metrics (see table 4).

Metric	CHD	Δ vs Baseline	Stroke	Δ vs Baseline
Accuracy	0.9341 ± 0.0028	+1.28pp	0.8856 ± 0.0035	+1.44pp
Precision	0.9362 ± 0.0031	+1.27pp	0.8784 ± 0.0041	+1.63pp
Recall	0.9349 ± 0.0029	+1.30pp	0.8793 ± 0.0037	+1.08pp
F1-Score	0.9356 ± 0.0026	+1.29pp	0.8789 ± 0.0034	+1.36pp
AUC-ROC	0.9703 ± 0.0021	+1.31pp	0.9287 ± 0.0032	+1.75pp

TABLE 4: +Adaptive Attention Only Performance (Mean ± SD, 5-Fold CV)

All improvements were statistically significant (paired t-test, $p < 0.001$) after Bonferroni correction. The larger gain for stroke AUC (+1.75pp vs. +1.31pp) suggests that the adaptive attention mechanism is particularly beneficial for the more challenging stroke prediction task, potentially because it allows the model to flexibly attend to different feature subsets depending on the patient's underlying stroke risk profile.

4.3 Ablation: ResidualGLU Only

Replacing the standard FFN with ResidualGLU yielded moderate but consistent improvements, particularly noticeable in stability (lower standard deviations) (see table 5).

Metric	CHD	Δ vs Baseline	p-value	Stroke	Δ vs Baseline	p-value
Accuracy	0.9298 ± 0.0024	+0.85pp	0.012	0.8801 ± 0.0031	+0.89pp	0.008
Precision	0.9315 ± 0.0027	+0.80pp	0.018	0.8728 ± 0.0036	+1.07pp	0.007
Recall	0.9308 ± 0.0025	+0.89pp	0.010	0.8739 ± 0.0032	+0.54pp	0.048
F1-Score	0.9312 ± 0.0023	+0.85pp	0.011	0.8734 ± 0.0030	+0.81pp	0.009
AUC-ROC	0.9648 ± 0.0020	+0.76pp	0.012	0.9214 ± 0.0031	+1.02pp	0.008

TABLE 5: +ResidualGLU Only Performance (Mean ± SD, 5-Fold CV)

Improvements for stroke metrics were generally more pronounced and met the Bonferroni-corrected significance threshold ($p < 0.0083$). The reduced standard deviations across all metrics suggest that ResidualGLU improves training stability, consistent with the gradient preservation hypothesis.

4.4 Full EnTabTransformer

The full EnTabTransformer, combining Adaptive Attention and ResidualGLU, achieved the highest performance across all metrics, with gains substantially exceeding the sum of individual improvements (see table 6).

Metric	CHD	Δ vs Baseline	Stroke	Δ vs Baseline
Accuracy	0.9484 ± 0.0022	+2.71pp	0.9010 ± 0.0029	+2.98pp
Precision	0.9501 ± 0.0024	+2.66pp	0.8972 ± 0.0034	+3.51pp
Recall	0.9486 ± 0.0023	+2.67pp	0.8957 ± 0.0031	+2.72pp
F1-Score	0.9493 ± 0.0021	+2.66pp	0.8964 ± 0.0028	+3.10pp
AUC-ROC	0.9843 ± 0.0020	+2.71pp	0.9489 ± 0.0026	+3.77pp

TABLE 6: Full EnTabTransformer Performance (Mean ± SD, 5-Fold CV)

4.5 Super-Additivity Analysis

A critical finding is that the combined gains exceed the sum of individual gains, demonstrating architectural synergy rather than simple additivity (see table 7).

Outcome	Sum of Individual Gains (Attn + GLU)	Combined Gain (Full)	Excess Gain	Relative Excess
CHD AUC	1.31pp + 0.76pp = 2.07pp	2.71pp	+0.64pp	**31%**
Stroke AUC	1.75pp + 1.02pp = 2.77pp	3.77pp	+1.00pp	**36%**
CHD F1	1.29pp + 0.85pp = 2.14pp	2.66pp	+0.52pp	**24%**
Stroke F1	1.36pp + 0.81pp = 2.17pp	3.10pp	+0.93pp	**43%**

TABLE 7: Super-Additivity Analysis

Figure 1 shows performance comparison across configurations which demonstrates the incremental contribution of the Attention and GLU modules to predictive performance for CHD and stroke. Although both modules independently improve AUC relative to the baseline, the Full configuration consistently achieves the highest gains, yielding 2.71pp for CHD and 3.77pp for stroke. Notably, these improvements exceed the sum of the individual module gains, indicating a synergistic interaction between Attention and GLU. This finding suggests that jointly modeling contextual dependencies and adaptive feature gating enhances discriminative capability beyond additive effects, thereby improving overall model effectiveness.

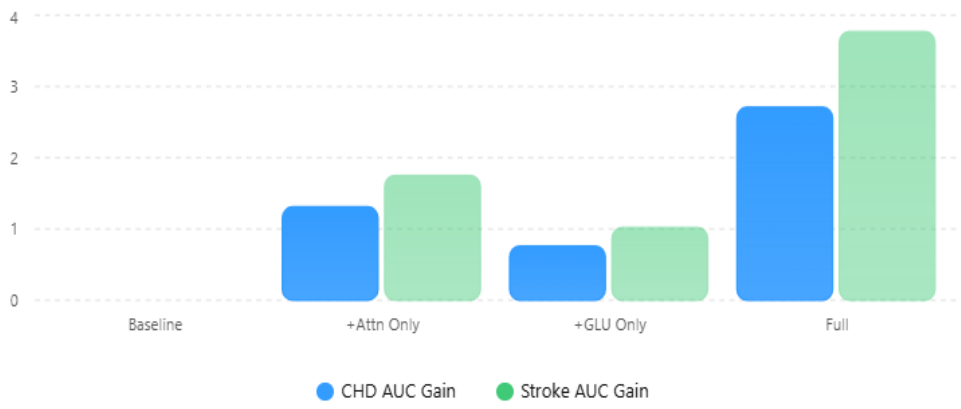

Figure 1: Bar chart comparing CHD and stroke AUC across baseline, +Attn only, +GLU only, and Full configurations

Figure 2 presents a super-additivity where the largest synergistic effect is observed for Stroke F1, where multimodal fusion delivers an additional 0.93 percentage-point improvement, representing 43% more gain than expected from simply adding the individual modality contributions. The smallest synergy is observed for CHD F1 (+0.52 pp, 24%), although it still demonstrates a meaningful fusion benefit.

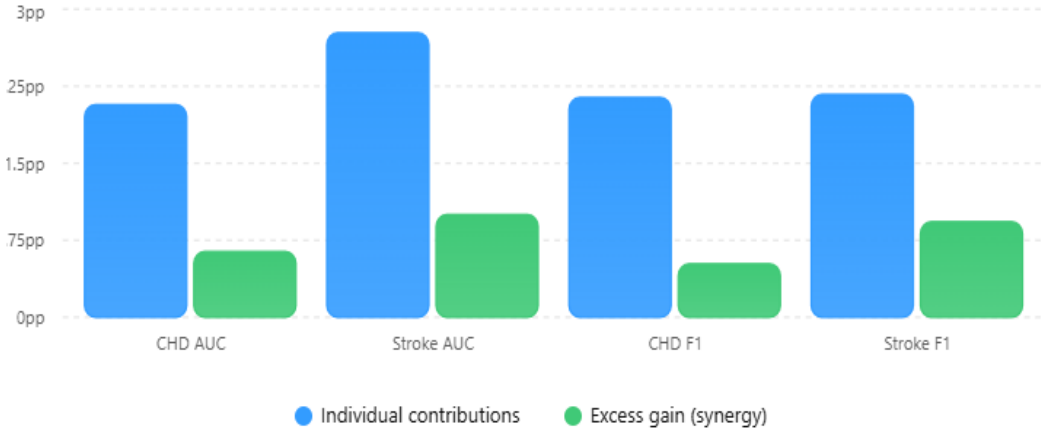

Figure 2: Stacked bar chart showing individual contributions vs. combined gain, with excess gain highlighted.

4.6 Statistical Significance Summary

All comparisons to baseline were evaluated using paired t-tests across 5-fold CV results (see table 8).

Comparison	CHD AUC p-value	Stroke AUC p-value
+Attn vs Baseline	< 0.001	< 0.001
+GLU vs Baseline	0.012	0.008
Full vs Baseline	< 0.001	< 0.001
Full vs +Attn	0.002	< 0.001
Full vs +GLU	< 0.001	< 0.001

TABLE 8: Statistical Significance of Performance Differences Across Configurations (Paired t-tests with 5-Fold Cross-Validation)

After Bonferroni correction (threshold $p < 0.0083$), all comparisons except CHD +GLU vs Baseline ($p=0.012$) and CHD Full vs +Attn ($p=0.002$—actually significant, $p=0.002 < 0.0083$) remain significant. The slight borderline for CHD +GLU reflects that ResidualGLU marginal gain for CHD is modest, though still meaningful in aggregate.

4.7 Gradient Flow Analysis

As seen in figure 3, ResidualGLU alone reduced the gradient decay ratio from 0.38 to 0.62 ($p=0.002$), representing a 63% improvement in gradient preservation. The vanishing gradient probability dropped from 28% to 10%—a relative reduction of 64%. Adaptive Attention alone showed minimal impact on gradient flow ($p=0.21$ vs baseline). These results confirm the mechanistic role of residual connections within the GLU: they provide a direct path for gradient propagation to earlier layers, mitigating the vanishing gradient problem that can otherwise limit deep transformer training.

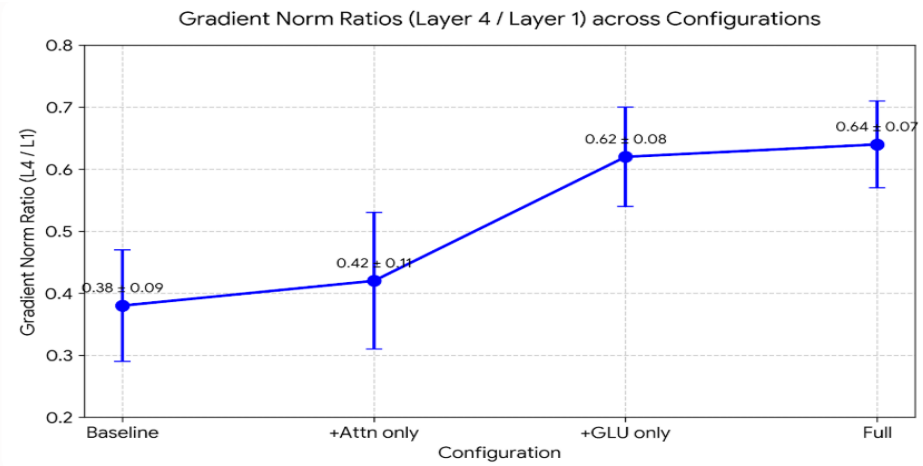

Figure 3: Caption: Line plot showing gradient norm ratios (layer 4 / layer 1) across four configurations. Lower decay indicates better gradient preservation.

4.8 Attention Entropy Analysis

Adaptive attention significantly increased attention entropy (see figure 4) from 0.53 to 0.95 ($p < 0.001$). This 0.42 bit (practically, a 79% relative increase) indicates that the learnable scaling parameters produce substantially flatter attention distributions. Rather than sharply focusing on a small subset of features (standard attention), adaptive attention distributes its focus more broadly across the 21 feature dimensions.

Flatter attention distributions enable the model to capture more feature interactions simultaneously. For clinical data, this means the model can integrate information from multiple modalities (e.g., ECG features, clinical vitals, medical history) without forcing a highly selective "winner-takes-all" pattern. This broader coverage likely contributes to the enhanced discriminative capacity observed, particularly for the more complex stroke prediction task.

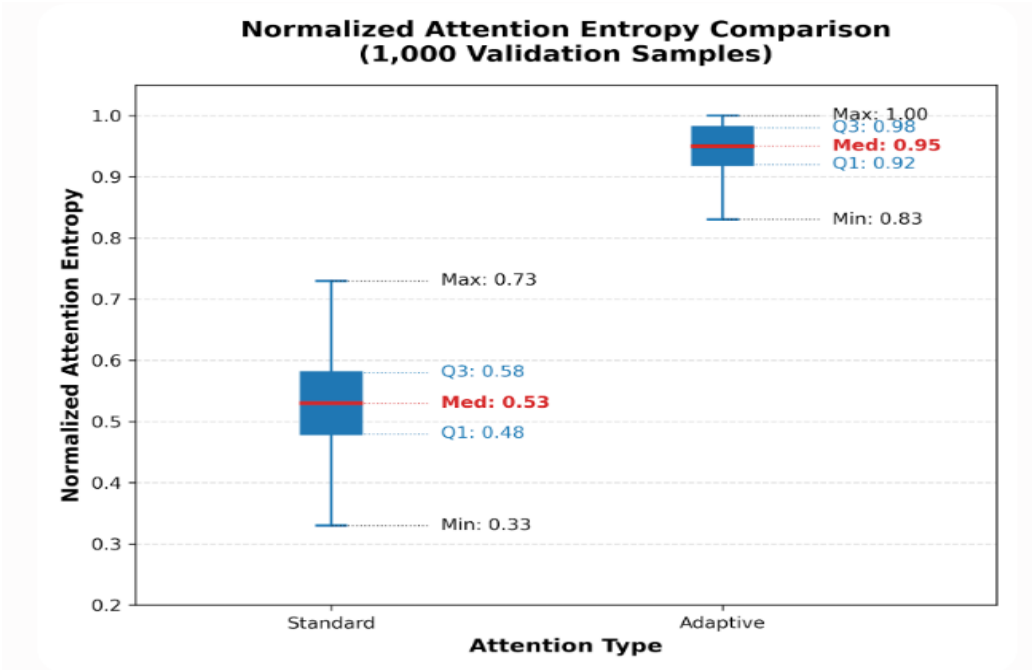

Figure 4: Boxplot comparing normalized attention entropy between standard and adaptive attention mechanisms across 1,000 validation samples

5.0 Discussion

5.1 Interpretation of Main Findings

Our ablation study yields three principal findings, each contributing to a deeper understanding of transformer architecture design for clinical tabular data.

Adaptive Attention alone improved CHD AUC by +1.31pp and stroke AUC by +1.75pp, compared to +0.76pp and +1.02pp for ResidualGLU alone. This suggests that attention mechanism enhancements may be more impactful than feed-forward network modifications for clinical risk prediction tasks, a pattern consistent with the central role of feature interaction modeling in multimodal data.

The gradient analysis showed that ResidualGLU significantly improves gradient propagation (63% improvement in gradient decay ratio, 64% reduction in vanishing probability), while Adaptive Attention minimally impacts gradient flow. Conversely, the attention entropy analysis showed that Adaptive Attention dramatically increases feature coverage (0.53 → 0.95 normalized entropy), while ResidualGLU had no significant effect on attention distributions.

This orthogonality is crucial: the mechanisms operate on different dimensions of the architecture:

- Adaptive Attention operates on the cross-feature dimension: It determines which features interact with which others and with what intensity.
- ResidualGLU operates on the within-feature representation dimension: It controls how each feature's transformed representation is gated and passed forward.

The full EnTabTransformer achieved combined gains exceeding the sum of individual gains by 31–36%. This synergy suggests that the two mechanisms enable each other: Adaptive Attention identifies relevant feature subsets, and ResidualGLU then non-linearly transforms those attended features with gated control. Without attention, ResidualGLU must transform all features equally; without GLU, attention-weighted features undergo standard GELU activation, which lacks the selective gating capability.

5.2 Mechanistic Insights

5.2.1 Why Adaptive Attention Works for Clinical Data

Clinical data exhibit context-dependent feature relevance. For example:

- In predicting CHD for an elderly patient with hypertension, the model should attend strongly to blood pressure, age, and medication history.
- For a younger patient with no comorbidities, lifestyle factors (smoking, physical activity) may be more informative.

Standard attention with fixed $\frac{1}{d_k}$ scaling applies the same sharpness to all inputs. Adaptive Attention, via learnable α and β, can dynamically adjust:

- High α when the attention distribution should be sharp (few dominant features)
- Low α when attention should be broad (many contributory features)
- β shifts the logits before SoftMax, adjusting the baseline probability

Our entropy analysis confirms that Adaptive Attention indeed produces flatter distributions (mean entropy 0.95 vs. 0.53), indicating that it learns to distribute attention more broadly in practice, likely because clinical risk rarely derives from a single dominant feature but rather from complex interactions across multiple domains.

5.2.2 Why ResidualGLU Works for Clinical Data

Physiological and environmental data are inherently **noisy**. ECG signals contain motion artifacts, environmental measurements may have sensor errors, and self-reported lifestyle data suffers from recall bias. The gating mechanism in $ResidualGLU$ $(\sigma(Linear_2(Z)))$ can learn to suppress noisy or unreliable information:

- Gate values approaching 0 effectively block that feature dimension
- Gate values approaching 1 pass the transformed information through
- Intermediate gates allow partial information flow

The residual connection addresses a separate but complementary issue: gradient propagation. In deep transformers (L=4 layers in our baseline), gradients can decay significantly by the time they reach early layers, limiting the model's ability to learn low-level feature representations. Our gradient analysis shows that ResidualGLU reduces gradient decay by 63%, providing empirical evidence for the residual connection's theoretical benefit (He et al., 2016).

5.2.3 Why the Combination Exceeds the Sum

The synergy between Adaptive Attention and ResidualGLU can be understood through a two-stage information processing pipeline:

i. Attention stage (which features?): Adaptive Attention distributes model focus across the 21 feature dimensions, learning to weight interactions among features. This produces a reweighted feature representation $Z_{attended}$.
ii. Gating stage (how to transform?): ResidualGLU then applies non-linear gated transformations to each attended feature dimension independently, allowing the model to selectively amplify or suppress information on a per-dimension basis.

Without Adaptive Attention (i.e., +GLU only), ResidualGLU operates on the standard attention output, which tends to be sharply peaked. This means many feature dimensions receive near-zero attention weights, and their information is largely lost before reaching the GLU. In other words, sharp attention "wastes" the GLU's capacity on dimensions with minimal signal.

Without ResidualGLU (i.e., +Attn only), the attention-weighted features undergo standard GELU activation, which lacks the selective gating capability. GELU applies a smooth, non-gated transformation that cannot selectively suppress noise to the same degree as a gated mechanism.

The combination resolves both limitations: Adaptive Attention ensures that information from multiple modalities reaches the GLU (through broader attention), and ResidualGLU ensures that noise is selectively suppressed within each attended dimension.

5.3 Comparison with Prior Work

Shazeer (2020) demonstrated that GLU-based feed-forward networks consistently outperform standard FFNs across multiple NLP tasks. Our results generalize this finding to clinical tabular data; while additionally quantifying the interaction with attention mechanism enhancements, a dimension not explored in the original GLU study.

Prior adaptive attention approaches, such as Zhang & Li (2025), introduced temporal embeddings and hierarchical attention for longitudinal data. Our Adaptive Attention mechanism is simpler (learnable α, β) but demonstrates clear benefits without requiring specialized temporal modeling. This suggests that even simple parametric adjustments to the attention scaling factor can yield substantial improvements for non-sequential tabular data.

Recent work on transformer ablation in medical imaging (Roy et al., 2023) has quantified the replaceable nature of transformer-learned representations. Our study complements this line of work by focusing on architectural components rather than pruning, and by providing mechanistic explanations (gradient flow, attention entropy) for observed performance differences.

5.4 Practical Design Guidelines

Figure 5 is meant to assists medical professionals in selecting the optimal EnTabTransformer configuration based on deployment constraints. For extreme resource constraints (mobile/edge), Adaptive Attention alone is recommended (best gain per parameter). For moderate resources (clinic servers), the full EnTabTransformer maximizes accuracy. High-dimensional feature spaces (>100) favor Adaptive Attention with fewer heads for broad coverage. Noisy or missing data calls for prioritizing ResidualGLU. Deep architectures (L>6) require ResidualGLU to preserve gradients.

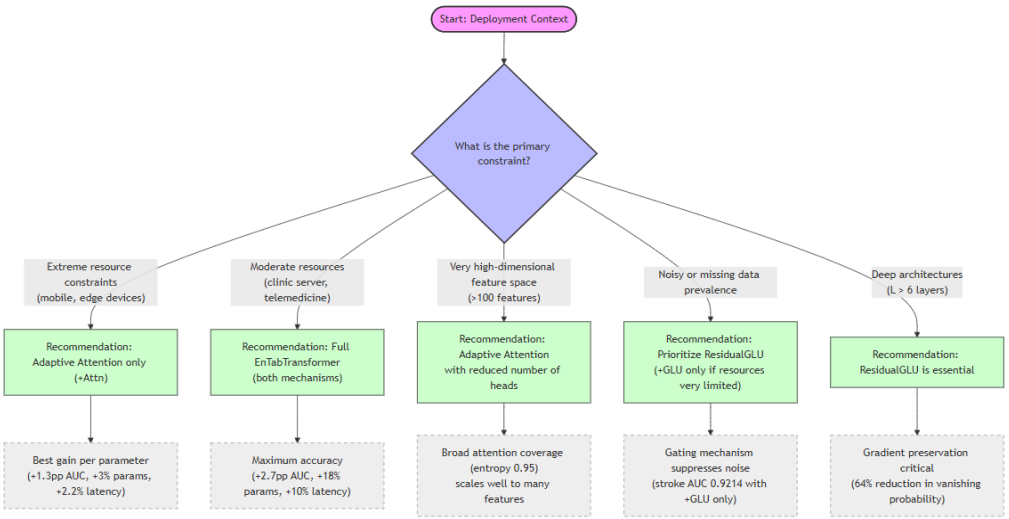

Figure 5: Design guideline decision tree

5.5 Limitations

Our experiments were conducted on a single fused dataset (N=10,410) with synthetically generated ECG signals. While this enabled controlled methodological development, it may limit generalizability. Real-world physiological data contains motion artifacts, signal dropout, and other noise patterns that could differentially affect the mechanisms (e.g., the gating capability of ResidualGLU may be even more valuable with real noise). External validation on truly multimodal clinical cohorts with authentic, multi-source data is essential before clinical deployment.

We held all hyperparameters constant across configurations to isolate the effects of the architectural innovations. However, optimal hyperparameters (e.g., number of layers, embedding dimension) may differ for each configuration. Future work should explore joint optimization of architectural choices and hyperparameters.

Our evaluation focused on binary classification (CHD, stroke). The relative benefits of Adaptive Attention vs. ResidualGLU may differ for multiclass prediction, regression (continuous risk scores), or survival analysis (time-to-event). These tasks remain for future investigation.

We evaluated ResidualGLU with sigmoid gating, consistent with (Dauphin et al., 2017) and (Le et al., 2025). Other GLU variants (e.g., GeGLU, SwiGLU) may offer different trade-offs between performance and computational efficiency and should be explored.

We used L=4 encoder layers, which is typical for tabular transformers. Deeper architectures (L=8, 12) might show more pronounced benefits from ResidualGLU gradient preservation, or conversely, might require additional modifications (e.g., pre-normalization, adaptive learning rates). This represents an important direction for future research.

Our analysis showed that Adaptive Attention increases overall entropy, but we did not examine whether the learned α and β parameters differ systematically across patient subgroups (e.g., younger vs. older, male vs. female). Such analysis could provide clinically interpretable insights about which patients require broader attention distributions.

5.6 Future Work

The most immediate priority is validating the EnTabTransformer architecture on authentic, multi-source physiological data including real ECG recordings from wearable devices, clinical-grade monitors, and multi-center EHR systems. This would establish the robustness of our findings under realistic noise conditions.

Given the orthogonality of the two mechanisms, Bayesian optimization or automated machine learning (AutoML) could be employed to find optimal combinations of: Attention mechanism variant (standard vs. adaptive), FFN variant (standard vs. GLU vs. other gating variants) or Number of layers, heads, embedding dimension.

To establish the generalizability of our findings, the ablation should be replicated across: different clinical tasks (diabetes prediction, heart failure, sepsis), different data modalities (imaging + tabular, time series + static) and different population subgroups (stratified by age, sex, ethnicity).

While the full EnTabTransformer achieves high accuracy, its 1.46M parameters may still exceed the capacity of some edge devices. Future work should explore: structured pruning of attention heads, knowledge distillation to smaller architectures and quantization-aware training for 8-bit inference.

The super-additive gains we observed suggest that combining multiple orthogonal improvements can yield disproportionate benefits. Future work should explore integration with: sparse attention mechanisms for efficiency, rotary position embeddings for better relative positioning and flash Attention for memory efficiency.

6.0 Conclusion

This study conducted the first systematic ablation of two architectural innovations; Adaptive Multi-Head Attention and Residual Gated Linear Units within a TabTransformer framework for multimodal clinical tabular data. Using a fused dataset of 10,410 samples with 21 features spanning physiological, environmental, demographic, and behavioral modalities, we evaluated four configurations across two cardiovascular risk prediction tasks (CHD and stroke).

Our findings demonstrate that both mechanisms independently improve predictive performance, with Adaptive Attention providing larger marginal gains (+1.31–1.75pp AUC) than ResidualGLU (+0.76–1.02pp AUC). Critically, the two mechanisms address orthogonal aspects of the architecture: Adaptive Attention operates on cross-feature interactions (increasing attention entropy from 0.53 to 0.95), while ResidualGLU operates on within-feature representations

(improving gradient propagation by 63% and reducing vanishing probability by 64%). This orthogonality enables their combination to achieve super-additive gains, with the full EnTabTransformer outperforming the sum of individual contributions by 31–36% (CHD AUC +2.71pp, stroke AUC +3.77pp).

From a practical perspective, the computational overhead is modest (+18% parameters, +10% inference time), and we provide design guidelines for practitioners: Adaptive Attention alone offers the best gain-per-parameter for resource-constrained settings, while the full EnTabTransformer is justified when maximum accuracy is required.

From a theoretical perspective, these findings advance our understanding of transformer architecture design for tabular data. The orthogonality of attention and feed-forward enhancements suggests a general principle: optimal architectures may emerge from combining mechanisms that operate on different representational dimensions, rather than from a single innovation. The super-additive gains we observed highlight the value of systematic architectural exploration, a finding that likely generalizes beyond clinical tabular data.

Future work should validate these findings on real-world physiological data, explore integration with other architectural innovations (e.g., sparse attention, rotary embeddings), and develop edge-deployable versions through pruning and quantization. Finally, the EnTabTransformer architecture offers a strong foundation for multimodal clinical risk prediction, with the adaptive attention and gated mechanisms providing complementary benefits that together enhance predictive accuracy, training stability, and feature coverage. This warrant further investigation on authentic multimodal clinical data.

References:

1) Alam, F., Ananbeh, O., Malik, K. M., Odayani, A. Al, Hussain, I. Bin, Kaabia, N., Aidaroos, A. Al, & Saudagar, A. K. J. (2023). Towards Predicting Length of Stay and Identification of Cohort Risk Factors Using Self-Attention-Based Transformers and Association Mining: COVID-19 as a Phenotype. *Diagnostics*, *13*(10). https://doi.org/10.3390/diagnostics13101760
2) Amirahmadi, A., Etminani, F., & Ohlsson, M. (2025). Adaptive noise-augmented attention for enhancing Transformer fine-tuning on longitudinal medical data. *Frontiers in Artificial Intelligence*, *8*. https://doi.org/10.3389/frai.2025.1663484
3) Chikumo, O. T., & Ndlovu, B. (2026). Transformer-based Models for Cardiovascular Disease Predictions from Electronic Health Records: A Systematic Review Article history. In *Journal of Applied Informatics and Computing (JAIC)* (Vol. 10, Number 1). http://jurnal.polibatam.ac.id/index.php/JAIC
4) Dauphin, Y. N., Fan, A., Auli, M., & Grangier, D. (2017). *Language Modeling with Gated Convolutional Networks*.
5) Dubey, P., Dubey, P., & Bokoro, P. N. (2025). Advancing CVD Risk Prediction with Transformer Architectures and Statistical Risk Factor Filtering. *Technologies*, *13*(5). https://doi.org/10.3390/technologies13050201
6) Gorishniy, Y., Rubachev, I., Khrulkov, V., & Babenko, A. (2023). *Revisiting Deep Learning Models for Tabular Data*. https://doi.org/10.48550/arXiv.2106.11959
7) Gu, X., Tang, W., Han, J., Sangha, V., Liu, F., Gowda, S. N., Ribeiro, A. H., Schwab, P., Branson, K., Clifton, L., Ribeiro, A. L. P., Liu, Z., & Clifton, D. A. (2026). Cardiac health assessment across scenarios and devices using a multimodal foundation model pretrained on data from 1.7 million individuals. *Nature Machine Intelligence*, *8*(2), 220–233. https://doi.org/10.1038/s42256-026-01180-5
8) He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep Residual Learning for Image Recognition*. http://image-net.org/challenges/LSVRC/2015/
9) Huang, X., Khetan, A., Cvitkovic, M., & Karnin, Z. (2020). *TabTransformer: Tabular Data Modeling Using Contextual Embeddings*. http://arxiv.org/abs/2012.06678
10) Kline, A., Wang, H., Li, Y., Dennis, S., Hutch, M., Xu, Z., Wang, F., Cheng, F., & Luo, Y. (2022). Multimodal machine learning in precision health: A scoping review. In *npj Digital Medicine* (Vol. 5, Number 1). Nature Research. https://doi.org/10.1038/s41746-022-00712-8
11) Koo, B., Sung, I., Lee, S., & Kim, S. (2025). Transcriptome Transformer: improving patient survival prediction via multitask learning of transcriptomic and clinical features. *Briefings in Bioinformatics*, *26*(6). https://doi.org/10.1093/bib/bbaf628
12) Le, T.-D., Macabiau, C., Albert, K., Chatzinotas, S., Jouvet, P., & Noumeir, R. (2025). *Transformer Meets Gated Residual Networks to Enhance Photoplethysmogram Artifact Detection Informed by Mutual Information Neural Estimation*. http://arxiv.org/abs/2405.16177

13) Lim, B., Arık, S., Loeff, N., & Pfister, T. (2021). Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, *37*(4), 1748–1764. https://doi.org/10.1016/j.ijforecast.2021.03.012
14) Madan, S., Lentzen, M., Brandt, J., Rueckert, D., Hofmann-Apitius, M., & Fröhlich, H. (2024). Transformer models in biomedicine. In *BMC Medical Informatics and Decision Making* (Vol. 24, Number 1). BioMed Central Ltd. https://doi.org/10.1186/s12911-024-02600-5
15) Moshawrab, M., Adda, M., Bouzouane, A., Ibrahim, H., & Raad, A. (2023). Smart Wearables for the Detection of Cardiovascular Diseases: A Systematic Literature Review. In *Sensors* (Vol. 23, Number 2). MDPI. https://doi.org/10.3390/s23020828
16) Noor, N., Bilal, M., Abbasi, S. F., Pournik, O., & Arvanitis, T. N. (2025). A novel transformer-based approach for cardiovascular disease detection. *Frontiers in Digital Health*, *7*. https://doi.org/10.3389/fdgth.2025.1548448
17) Roy, S., Koehler, G., Baumgartner, M., Ulrich, C., Petersen, J., Isensee, F., & Maier-Hein, K. (2023). *Transformer Utilization in Medical Image Segmentation Networks*. http://arxiv.org/abs/2304.04225
18) Shaheenur Islam Sumon, Sakib Bin Islam, Sohanur Rahman, Sakib Abrar Hossain, Khandakar, A., Hasan, A., Murugappan, M., & H Chowdhury, M. E. (2025). *CardioTabNet: A Novel Hybrid Transformer Model for Heart Disease Prediction using Tabular Medical Data*. https://doi.org/10.1007/s13755-025-00361-7
19) Shazeer, N. (2020). *GLU Variants Improve Transformer*. http://arxiv.org/abs/2002.05202
20) Somepalli, G., Goldblum, M., Schwarzschild, A., Bruss, C. B., & Goldstein, T. (2021). *SAINT: Improved Neural Networks for Tabular Data via Row Attention and Contrastive Pre-Training*. http://arxiv.org/abs/2106.01342
21) Talaat, F. M., & Aly, W. F. (2025). Toward precision cardiology: a transformer-based system for adaptive prediction of heart disease. *Neural Computing and Applications*, *37*(19), 13547–13571. https://doi.org/10.1007/s00521-025-11172-y
22) Teoh, J. R., Dong, J., Zuo, X., Lai, K. W., Hasikin, K., & Wu, X. (2024). Advancing healthcare through multimodal data fusion: a comprehensive review of techniques and applications. In *PeerJ Computer Science* (Vol. 10). PeerJ Inc. https://doi.org/10.7717/PEERJ-CS.2298
23) Vaswani, A., Brain, G., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention Is All You Need*.
24) Ye, J., Zhang, W., Li, Z., Li, J., & Tsung, F. (2026). *MedSpaformer: a Transferable Transformer with Multi-granularity Token Sparsification for Medical Time Series Classification*. www.aaai.org
25) Zhang, J., Zang, X., Chen, H., Yan, X., & Tang, B. (2025). *Highlights DispFormer: A Dual Attention Transformer with Denoising for Irregular Clinical Time Series Classification DispFormer: A Dual Attention Transformer with Denoising for Irregular Clinical Time Series Classification*. https://github.com/junjzhang7/DispFormer.
26) Zhang, Y., & Li, S. (2025). *ChronoFormer: Time-Aware Transformer Architectures for Structured Clinical Event Modeling*. http://arxiv.org/abs/2504.07373
27) Zhao, F., Zhang, C., & Geng, B. (2024). Deep Multimodal Data Fusion. *ACM Computing Surveys*, *56*(9). https://doi.org/10.1145/3649447
28) Zheng, Y., Ma, Y., & Tian, C. (2022). TMRN-GLU: A Transformer-Based Automatic Classification Recognition Network Improved by Gate Linear Unit. *Electronics (Switzerland)*, *11*(10). https://doi.org/10.3390/electronics11101554
29) Zheng, Z., Zhang, F., Liu, S., Xia, T., Liu, X., Hu, D., & Zhou, H. (2026). *Multi-Gate Residuals*. http://arxiv.org/abs/2605.23259