

ORIGINAL RESEARCH ARTICLE

Comparative study of transformer-based models for cardiovascular disease risk stratification with tabular biomarker data

Yuvraj Sharma^{1,2}, **Ekta Tiwari³**, **Neha Gupta²**, **Mustafa Al-Maini⁴**, **Rajesh Singh⁵**, **Narendra N. Khanna⁶**, **Vijay Rathore⁷**, **Vandana Kumari⁸**, **John R. Laird⁹**, **George Kitas^{10,11}**, **Andrew Nicolaides¹²**, **Zoltan Ruzsa¹³**, **Aditya Sharma¹⁴**, **Amer M. Johri¹⁵**, **Laura E. Mantella¹⁶**, **Gavino Faa^{17,18}**, **Esma R. Isenovic¹⁹**, **Luca Saba²⁰**, **Jasjit S. Suri^{1,21,22,23*}**

¹Stroke Diagnostic and Monitoring Division, AtheroPoint™, Roseville, California, United States of America

²Department of Information Technology, Bharati Vidyapeeth's College of Engineering, New Delhi, India

³Department of Innovation, Global Biomedical Technologies, Inc., Roseville, California, United States of America

⁴Department of Allergy, Clinical Immunology and Rheumatology Institute, Toronto, Ontario, Canada

⁵Department of Research and Innovation, Uttaranchal Institute of Technology, Uttaranchal University, Dehradun, Uttarakhand, India

⁶Department of Cardiology, Apollo Hospitals, New Delhi, India

⁷Department of Nephrology, Kaiser Permanente, Sacramento, California, United States of America

⁸School of Computer Science and Engineering, Galgotias University, Greater Noida, Uttar Pradesh, India

⁹Department of Cardiology, St. Helena Hospital, St. Helena, California, United States of America

¹⁰Academic Affairs, Dudley Group National Health Service Foundation Trust, Dudley, West Midlands, United Kingdom

¹¹Arthritis Research United Kingdom Epidemiology Unit, Manchester University, Manchester, Greater Manchester, United Kingdom

¹²Vascular Screening and Diagnostic Centre, University of Nicosia, Nicosia, Cyprus

¹³Invasive Cardiology Division, University of Szeged, Szeged, Hungary.

¹⁴Division of Cardiovascular Medicine, University of Virginia, Charlottesville, Virginia, United States of America

¹⁵Department of Medicine, Queen's University, Kingston, Ontario, Canada

¹⁶Department of Medicine, University of Toronto, Toronto, Ontario, Canada

¹⁷Department of Biology, College of Science and Technology, Temple University, Philadelphia, Pennsylvania, United States of America

¹⁸Department of Medical Sciences and Public Health, University of Cagliari, Cagliari, Sardinia, Italy

¹⁹Department of Radiobiology and Molecular Genetics, National Institute of the Republic of Serbia, University of Belgrade, Belgrade, Serbia

²⁰Department of Radiology, Azienda Ospedaliero Universitaria, Cagliari, Sardinia, Italy

²¹University Center for Research and Development, Chandigarh University, Mohali, Punjab, India

²²Department of Computer Science, Symbiosis Institute of Technology, Symbiosis International (Deemed University), Nagpur, Maharashtra, India

²³Department of Electrical and Computer Engineering, Idaho State University, Pocatello, Idaho, United States of America

*Corresponding author:

Jasjit S. Suri
(jasjit.suri@atheropoint.com)

Citation: Sharma Y, Tiwari E, Gupta N. Comparative study of transformer-based models for cardiovascular disease risk stratification with tabular biomarker data. *Artif Intell Health*. 2026;3(3):025460100. doi: 10.36922/AIH025460100

Received: November 14, 2025

Revised: December 19, 2025

Accepted: January 8, 2026

Published online: February 4, 2026

Copyright: © 2026 Author(s). This is an Open-Access article distributed under the terms of the Creative Commons Attribution License, permitting distribution, and reproduction in any medium, provided the original work is properly cited.

Publisher's Note: AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Abstract

Cardiovascular disease (CVD) remains one of the leading causes of global mortality and disability. Advances in computational modeling and artificial intelligence have enhanced CVD risk prediction by integrating multivariate clinical and biochemical features. Recent developments in deep learning, especially transformer-based models for tabular data, have demonstrated superior capabilities in capturing nonlinear and high-dimensional biomarker interactions. This study proposes a predictive framework that uses statistical and clinical biomarker data to assess CVD risk. Traditional machine learning models (logistic regression, Gaussian Naïve Bayes, linear discriminant analysis, AdaBoost, and XGBoost) were compared with deep learning models (gated recurrent unit [GRU] and long short-term memory [LSTM]) and transformer-based models—self-attention and intersample attention transformer (SAINT), feature tokenizer (FT), and tab transformer. Experiments were conducted to evaluate the impact of data augmentation, analyze learning behavior through loss and accuracy curves, and assess a fusion approach combining tab transformer with recurrent networks. Model performance was evaluated using accuracy and receiver operating characteristic analysis. Transformer-based models consistently outperformed conventional machine learning and deep learning methods. SAINT and FT achieved an area under the curve (AUC) of 0.8875 and 0.9489, respectively. The tab transformer demonstrated the highest performance with an AUC of 0.9728. The fusion of the tab transformer with GRU and LSTM further enhanced predictive precision, improving representation learning and generalization for CVD risk prediction. The proposed transformer-based framework offers a robust, scalable, and interpretable solution for accurate CVD risk assessment. Its superior predictive capability highlights the potential for integration into clinical decision-support systems for early diagnosis and patient management.

Keywords: Cardiovascular disease; Deep learning; Transformer; Tab transformer; Machine learning; Risk prediction; Biomedical data

1. Introduction

Cardiovascular Disease (CVD) remains a major global cause of death, particularly among patients with diabetes mellitus¹ and chronic kidney disease.² Various biomarkers have been identified to monitor disease progression, including indicators of neurological dysfunction, blood viscosity, and hypertension. Additional factors—such as smoking history, radiation exposure, and metabolic disorders—further elevate cardiovascular risk. This study focuses exclusively on clinical biomarker data for disease detection, emphasizing statistical and biochemical features over imaging-based diagnostics.

Transformer-based models for tabular data—feature tokenizer (FT),³ self-attention and intersample attention transformer (SAINT),⁴ and tab transformer⁵—are introduced to enhance predictive accuracy in CVD assessment. These models leverage attention mechanisms to capture complex feature interactions that traditional

methods often miss. FT tokenizes features for effective embedding; SAINT incorporates self- and inter-sample attention for richer relational learning; and tab transformer models both categorical and continuous features using context-aware embeddings, improving interpretability and generalization across patient data. This work is a benchmarking framework for CVD risk prediction that evaluates classical machine learning (ML) models, deep learning (DL) models (gated recurrent unit [GRU] and long short-term memory [LSTM]),^{6–9} and three advanced transformer-based models (SAINT, FT, tab transformer), under a standardized pipeline.

The experimental framework evaluates model performance through sequential analyses, starting with the effect of data augmentation on generalization and stability. Traditional ML models (logistic regression, Gaussian Naïve Bayes, linear discriminant analysis (LDA), AdaBoost, and XGBoost),^{10–14} DL models (GRU, LSTM),^{15–17} and transformer-based models (FT, SAINT,

tab transformer) are compared using accuracy, loss, and receiver operating characteristic (ROC) metrics. A fusion approach combining the tab transformer with recurrent models further enhances predictive precision and overall robustness.

The proposed system in Figure 1 integrates preprocessing, feature extraction, and model training in a scalable pipeline. Clinical data are normalized, balanced, and augmented before being processed using transformer-based models for feature learning and classification. As shown in Figure 1, this framework supports robust, automated, and clinically adaptable prediction of CVD risk.

The remainder of the current paper presents background literature, followed by biomarker characterization, data preprocessing, experimental protocols, and model evaluations. The study concludes with key findings, performance outcomes, and future prospects for transformer-based risk stratification.

2. Background literature

CVD risk stratification has evolved over decades, progressing from traditional statistical models to advanced DL frameworks. Early prediction approaches in the 1980s–1990s relied on demographic and clinical variables, such as gender, age, cholesterol level, blood pressure, and smoking history.¹⁸ During the 2000s, multivariable models incorporated additional parameters, including diabetes, family history, and body mass index (BMI), alongside region-specific Systematic Coronary Risk Evaluation charts.¹⁹ The 2010s marked the emergence of ML techniques—such as random forests, support vector machines, and neural networks—integrated with broader clinical and therapeutic features.²⁰ AtheroEdge™ systems and Detrano *et al.*²¹ validated that incorporating imaging biomarkers, such as carotid intima-media thickness, total plaque area (TPA), and coronary calcium scores, further

enhanced diagnostic accuracy. Since 2015, DL methods using convolutional neural networks and recurrent neural networks have advanced CVD risk prediction by identifying complex patterns within electronic health records and retinal imagery.²² Recently, Transformer-based models—FT, SAINT, and tab transformer—have achieved state-of-the-art performance on tabular biomarker data, effectively modeling nonlinear relationships and improving both predictive accuracy and interpretability in CVD assessment.

3. Methodology

This study was conducted using data gathered from one clinical setting, the Cardiovascular Imaging Network at Queen's University, Canada, and was later provided to AtheroPoint™, USA, under institutional agreements. This dataset is not publicly accessible and consists of two independent cohorts of patients: an experimental dataset of 500 patients and a validation dataset of 459 patients, representing real-world populations for cardiovascular risk assessment. Every patient record has a systematic collection of 21 cardiovascular biomarkers, divided into office-based, laboratory-based, carotid ultrasound image-based parameters, and MedUse groups. These are demographic and clinical factors, including age, sex, diabetes, hypertension, hyperlipidemia, smoking history, BMI, creatinine, and glomerular filtration rate (GFR). Other variables include family history, presence of angina, and the use of medications (statins, angiotensin-converting enzyme inhibitors, angiotensin 2 receptor blockers, beta-blockers). Radiomic parameters derived from ultrasound, such as TPA, minimal plaque height, and intra-plaque neovascularization,^{23–25} were also incorporated, and AngioScore (0 to 1) was the target label.

The research sample was representative of a mixed cardiovascular prevention cohort, comprising individuals

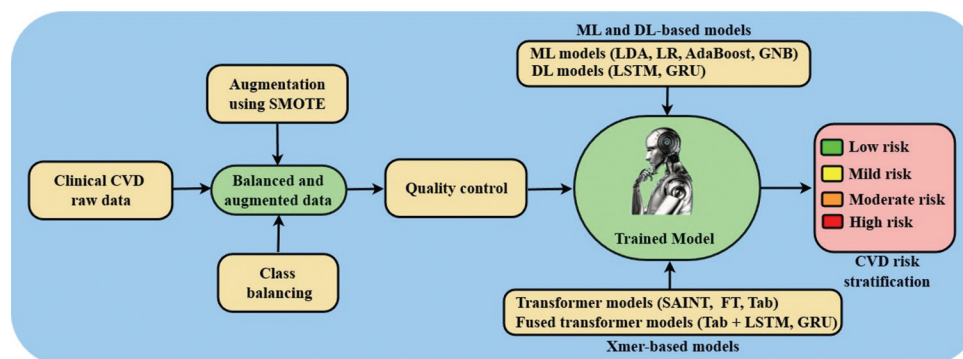


Figure 1. Online system design

Abbreviations: CVD: Cardiovascular disease; DL: Deep learning; FT: Feature tokenizer; GNB: Gaussian Naïve Bayes; GRU: Gated recurrent unit; LDA: Linear discriminant analysis; LR: Logistic regression; LSTM: Long short-term memory; ML: Machine learning; SAINT: Self-attention and intersample attention transformer; SMOTE: Synthetic minority oversampling technique.

who underwent vascular screening, those with known risk factors, and those with subclinical atherosclerosis detected using ultrasound. We provided statistically significant demographic data, including age range, mean age, and sex distribution. Since the dataset was skewed and had fewer CVD-positive cases, we applied the synthetic minority oversampling technique (SMOTE)^{26,27} to the training folds to prevent leakage. To obtain a balanced dataset of 3,060 records (765 per class), we applied SMOTE and controlled random sampling. This approach enhanced feature uniformity and the model's resilience while ensuring that only real patient data were included in the validation and test sets.

Each experiment was performed using a 10-fold stratified cross-validation protocol, with each cohort included. The splits were performed at the patient level, and no patient appeared in both the training and validation folds. Temporal biomarker alignment was performed among the patients alone, avoiding temporal or identity leakage. Figure 2 illustrates the experimental workflow used to assess three transformer-based models (FT, SAINT, and tab transformer) with recurrent layers to create our proposed fusion models. A transformer block was used in this design, in which the cardiovascular biomarker inputs were first converted into contextual embeddings, and the resulting representation was then fed into a GRU/LSTM unit to extract the temporal and sequential patterns.

The fused output was then processed by a dense layer and optimized using the Adam optimizer. This parallel-transformer-plus-recurrent structure enabled the model to learn both cross-feature relationships and longitudinal trends, producing more informative representations than a standalone transformer. Compared to the tab transformer alone, the fusion version achieved a measurable improvement in predictive accuracy with a modest increase in computational cost due to the additional recurrent parameters. Training time increased slightly, but inference remained efficient and suitable for

clinical deployment. This model demonstrated enhanced robustness and generalization for CVD risk prediction.

3.1. Self-attention and intersample attention transformer

The SAINT model uses a dual-attention model to learn feature-level and sample-level dependencies in tabular data. It has two modules, as illustrated in Figure 3, a self-attention block and an intersample attention block. The self-attention block is trained with multi-head attention, normalization, and feedforward layers to learn intra-sample feature relationships, whereas the intersample block is trained to learn cross-sample dependencies, which strengthen the contextual representation. Such a two-phase design allows SAINT to capture complex biomedical interactions, thereby improving its ability to predict CVD risk.

3.2. Feature tokenizer transformer

The FT transformer is a modification of the transformer framework for tabular data, in which each feature is represented as a token, as shown in Figure 4. It simultaneously works with both categorical and continuous variables in a single embedding space, eliminating the requirement of heavy preprocessing. These embeddings are processed by regular transformer encoder layers of multi-head attention and feedforward networks. FT learns inter-feature dependencies and nonlinear correlations by treating features as tokens, making it an effective way to generalize effectively on small medical data. FT demonstrated a remarkable predictive power in CVD risk assessment in this study.

3.3. Tab transformer

The tab transformer uses context-aware embeddings to encode categorical and continuous features, which substitutes one-hot encodings with transformer-encoded representations. This enables the dynamically adaptable contextual feature relationship and higher-

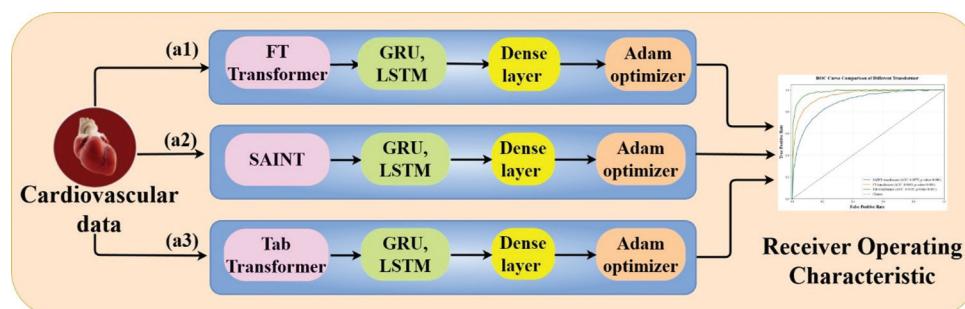


Figure 2. Tandem block diagram

Abbreviations: FT: Feature Tokenizer; GRU: Gated recurrent unit; LSTM: Long short-term memory; SAINT: Self-attention and intersample attention transformer; TAB Xmer: Tab transformer; Xmer: Transformer.

order dependencies across heterogeneous data. The tab transformer exhibited the highest area under the curve (AUC) among all models for CVD risk prediction because it better models feature interactions and is easier to interpret, as evidenced in Figure 5. It is strong and explainable, which makes it suitable for clinical use where precise risk stratification based on biomarkers is needed.

3.4. Experimental protocols

The experimental framework was designed to evaluate the reliability, interpretability, and predictive performance of transformer-based models—SAINT, tab transformer, and FT—in comparison with conventional DL and ML techniques for CVD risk prediction. The setup focused on assessing the effects of data augmentation, model

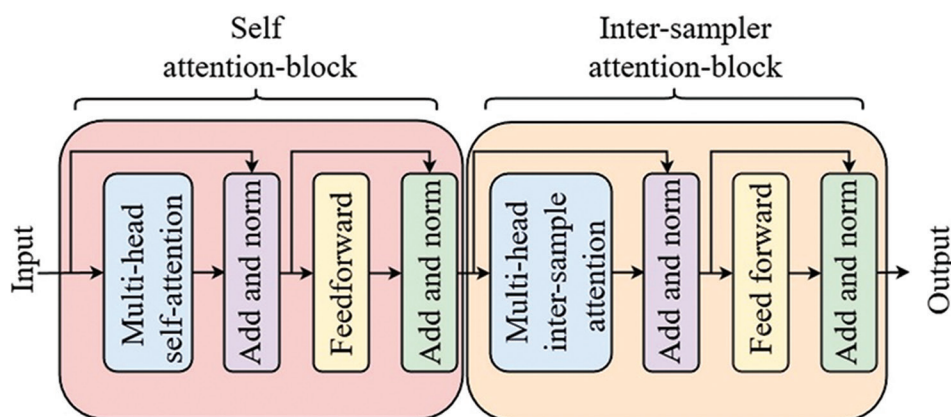


Figure 3. Cell architecture of self-attention and intersample attention transformer

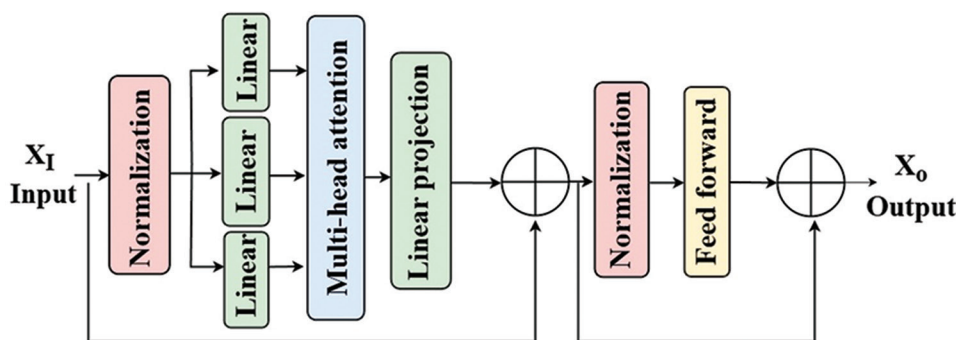


Figure 4. Cell architecture of feature tokenizer transformer

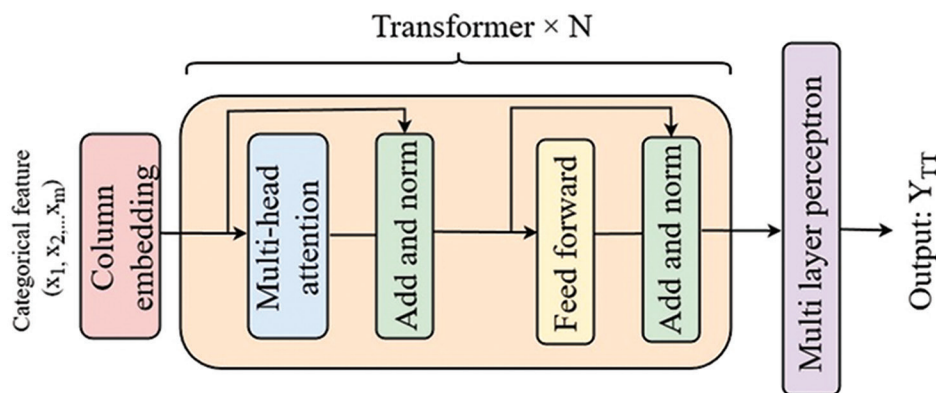


Figure 5. Cell architecture of the tab transformer

stability, and feature-learning capability within a unified experimental environment.

3.4.1. Experiment 1: Effect of augmentation

The first experiment concerned the impact of data augmentation on the model's generalization and learning performance. Synthetic balancing increased the dataset, improving the representation of minority classes and stabilizing the learning dynamics. This operation enhanced predictive consistency and strength across transformer structures.

3.4.2. Experiment 2: Model comparison (machine learning vs. deep learning vs. transformers)

The second experiment involved a three-model group comparison (traditional ML [logistic regression, AdaBoost, Gaussian Naïve Bayes, LDA, XGBoost], DL [GRU, LSTM], and transformer-based model [SAINT, tab transformer, FT]) comparison. Results show that transformers learned interactions more effectively, with superior generalization and outcomes, than classical and deep models in the association among intricate biomarkers.

3.4.3. Experiment 3: Accuracy and loss curves

Transformer-based models showed convergence and stability in the learning curves. The gradual decrease in loss and improvement in accuracy per epoch indicated effective optimization and minimal overfitting, demonstrating that transformers could effectively encode complex biomarker dependencies.

3.4.4. Experiment 4: Fusion effect

The final experiment was a hybrid fusion approach that integrated the most efficient tab transformer and sequential models (LSTM, GRU). These features combined the contextual representations of the tab transformer and the temporal modeling of recurrent networks to achieve enhanced feature representation and accurate CVD risk prediction.

3.5. Optimization

The system used grid-search cross-validation to estimate the most efficient hyperparameter settings to maximize model performance. A large number of parameter tuning runs were performed to balance training stability and predictive accuracy across transformer-based, DL, and ML models. The best setup was the learning rate of 0.00007, Adam optimizer, softmax activation, a batch size of 32, and 200 training steps. Other parameters were also optimized, including a dropout rate of 0.2 and a weight decay of 0.001, to avoid overfitting and enhance convergence. To ensure a fair comparison, all deep and transformer-based models

were trained using the same two-stage tuning scheme, in which the initial step was a coarse search, followed by a refinement to the most promising settings. The FT was also a feature of transformer-based models, in which categorical variables were transformed into learnable embeddings and numerical variables were transformed into linear projections, allowing the model to process all features in a single embedding space. These design decisions enabled the transformer-based models, specifically the tab transformer, to exhibit stable learning behavior and to provide better predictive performance than traditional methods for CVD risk stratification.

4. Results

We assessed the suggested framework across several aspects, including data augmentation, learning behavior, model comparison, and statistical validation. All experiments indicated that predictive performance depended on various factors, including data variety and model combination. The tab transformer showed greater accuracy and stability in CVD risk prediction.

4.1. Results of experiment 1: Effect of augmentation

Figure 6 shows the impact of data augmentation on model accuracy. The baseline model achieved 69.32% accuracy, improving to 76.20% and 83.95% with 2 × and 3 × augmentation, respectively. Accuracy peaked at 89.98% for 9 × augmentation before slightly decreasing to 88.36% at 10 ×. Moderate augmentation (4–9 ×) enhanced generalization and stability, while excessive augmentation yielded diminishing returns.

4.2. Results of experiment 2: Model comparison

Table 1 summarizes the performance of ML, recurrent, and transformer-based models based on 10-fold cross-validation. Older models (Gaussian Naïve Bayes, logistic regression, LDA, AdaBoost, XGBoost) exhibited moderate accuracy (53–86%) and AUC (0.67–0.91). GRU and LSTM

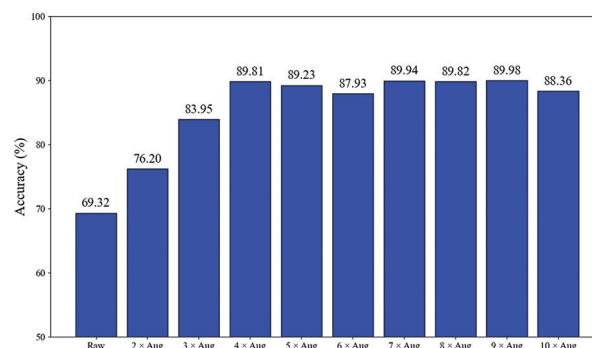


Figure 6. Effect of augmentation (Aug) using the 10-fold cross-validation protocol

Table 1. Performance comparison of different models using the 10-fold cross-validation protocol

Models	Accuracy (%)	Precision (%)	Sensitivity (%)	F1 (%)	AUC [0–1]
GNB	53.24	53.33	58.24	54.92	0.6703
Logistic regression	56.34	56.38	56.34	56.26	0.7881
LDA	59.61	59.73	59.61	59.39	0.8137
AdaBoost	66.23	65.89	66.32	65.61	0.8169
XGBoost	86.31	87.24	87.32	87.53	0.9178
GRU	87.04	87.12	87.21	87.23	0.9134
LSTM	87.86	87.89	87.65	87.56	0.9257
SAINT	86.16	86.45	86.78	86.64	0.8875
FT transformer	89.18	89.37	89.78	89.76	0.9489
Tab transformer	95.18	95.72	95.78	95.42	0.9728

Abbreviations: AUC: Area under the curve; FT: Feature tokenizer; GNB: Gaussian Naïve Bayes; GRU: Gated recurrent unit; LDA: Linear discriminant analysis; LR: Logistic regression; LSTM: Long short-term memory; SAINT: Self-attention and intersample attention transformer.

recurrent models performed better, with accuracies over 87% and AUCs over 0.91. Transformer-based models performed remarkably, with accuracies of 86.16%, 89.18%, and 95.18%, and AUCs of 0.8875, 0.9489, and 0.9728 for SAINT, FT transformer, and tab transformer, respectively, suggesting their superior efficacy in predicting complex interactions between biomarkers.

4.3. Results of experiment 3: Loss and accuracy curves

Figures 7 and 8 depict the training and validation performance of the tab transformer. Both loss curves declined smoothly over 200 epochs, indicating stable convergence. The close alignment between training and validation losses suggested minimal overfitting and strong generalization. Accuracy steadily improved, exceeding 90% in later epochs, confirming the model's robustness for CVD prediction.

4.4. Results of experiment 4: Fusion effect

4.4.1. Performance evaluation

The ROC curve in Figure 9 compares SAINT, FT, and the tab transformer for CVD risk prediction. All models performed robustly, with the tab transformer showing the best predictive accuracy (AUC = 0.9728), followed by FT (0.9489) and SAINT (0.8875). Their high AUCs suggested superior ability to distinguish positive from negative CVD cases, demonstrating strong predictive power for clinical data.

The transformer-based models were fused with DL models (GRU and LSTM), and the fusion models' predictive performance was assessed using ROC analysis. With recurrent networks added, the models' performance

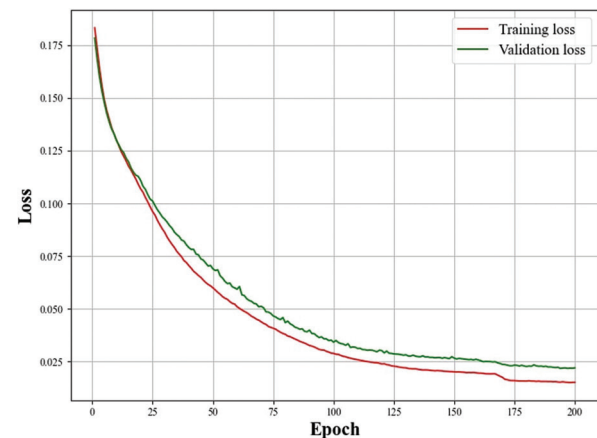


Figure 7. Train versus validation loss curve for the tab transformer, limiting to 200 epochs

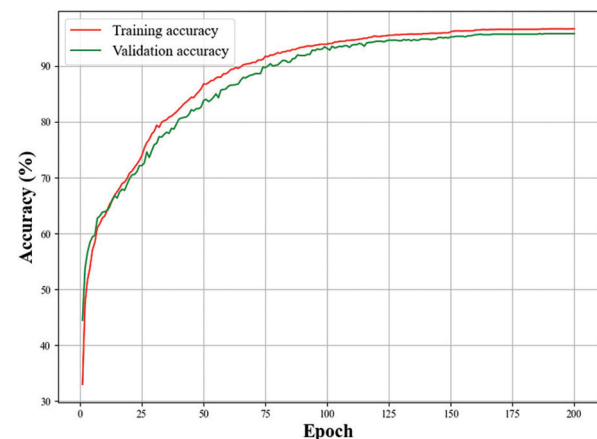


Figure 8. Train versus validation accuracy curve for the tab transformer, limiting to 200 epochs

further improved: GRU-tab and LSTM-tab demonstrated AUCs of 0.9823 and 0.9829, respectively, indicating improved temporal feature extraction and better ability to predict high-risk patients compared to the tab transformer alone (AUC: 0.9728). In contrast, the FT and SAINT fusion models (GRU-FT, LSTM-FT, GRU-SAINT, and LSTM-SAINT) demonstrated smaller improvements than those of the tab-based models as depicted in Figure 9 (ROC analysis), while the corresponding fusion accuracy comparison is shown in Figure 10.

4.4.2. Power analysis

Using the augmented dataset of 2,401 patients yielded statistically reliable and generalizable results. The observed effect sizes and dataset scale confirm the tab transformer's

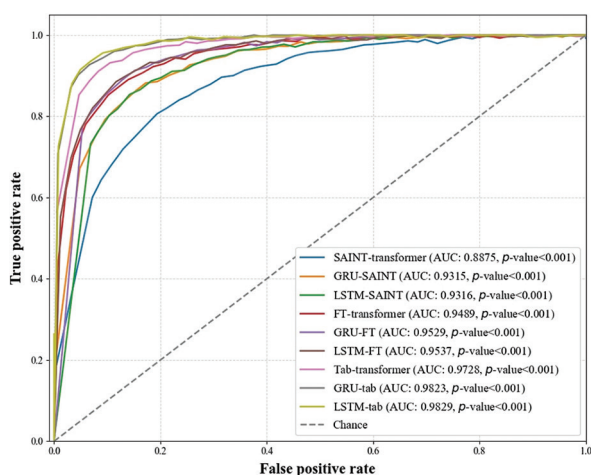


Figure 9. Receiver operating characteristic curves of different transformer-based models

Abbreviations: AUC: Area under the curve; FT: Feature tokenizer; GRU: Gated recurrent unit; LSTM: Long short-term memory; SAINT: Self-attention and intersample attention transformer.

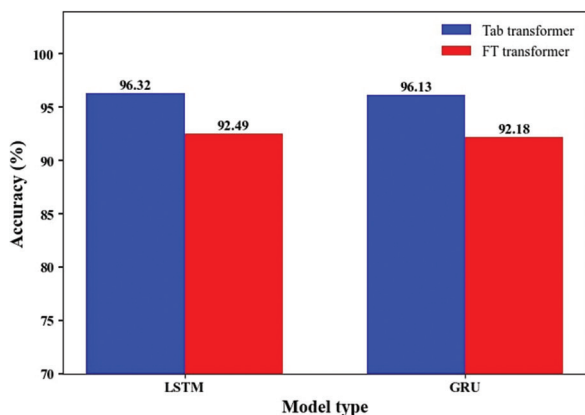


Figure 10. Accuracy comparison of the fusion models

effectiveness and superiority over traditional methods for CVD risk prediction, as shown in Equation (1):

$$N^* = \frac{z^2 \times p \times (1-p)}{MoE^2} \quad (1)$$

Where N^* represents the minimum required sample size, $z = 1.96$ denotes the Z-score corresponding to the desired confidence level, $p = 0.5$ (worst case) is the assumed proportion or variability, and $MoE = 0.02$ stands for the margin of error.

5. Scientific and clinical validation

To ensure the clinical reliability of the proposed framework, we performed a two-part scientific validation focusing on interpretability, statistical robustness, and clinical utility. Shapley additive explanation (SHAP) analysis was used to confirm that the model's key predictors align with established cardiovascular risk factors, while multi-seed evaluations and multiple significance tests verified the stability and reproducibility of model performance.

5.1. Explainability

To assess the interpretability of the proposed framework, we performed a SHAP analysis to determine each biomarker's contribution to the model's predictions. The SHAP

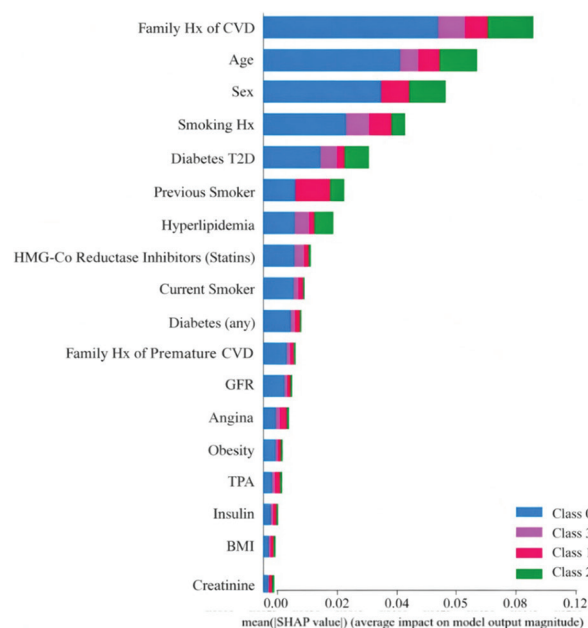


Figure 11. Class-wise SHAP values for tab transformer

Abbreviations: BMI: Body Mass Index; CVD: Cardiovascular disease; GFR: Glomerular Filtration Rate; Hx: History; IPN: Intravascular Neovascularization; SHAP: Shapley additive explanation; TPA: Total Plaque Area.

summary findings (Figure 11; Class 0 = low risk, Class 1 = mild risk, Class 2 = moderate risk, and Class 3 = high risk) indicate that family history of CVD, age, sex, smoking history, type 2 diabetes, and hyperlipidemia were the most significant predictors of CVD risk.

These variables were the most strongly related to model output across all classes and aligned with well-known clinical risk factors reported in cardiovascular epidemiology. There was a moderate effect on medication-related characteristics, such as statin use and other metabolic indicators, including GFR and creatinine, and smaller but significant effects on radiomic characteristics, including TPA and BMI. The correspondence between rankings of SHAP-derived importance and established CVD determinants was a strong indication that the model represented physiologically plausible relationships rather than relying on spurious statistical artifacts. This interpretability analysis confirmed the clinical credibility of the proposed framework and justified its possible application in the risk stratification context.

5.2. Statistical test

To ensure the strength and scientific validity of the proposed models, we conducted a comprehensive set of statistical tests to validate the multi-seed evaluations and the significance tests. Transformer-based and DL models were retrained with three independent random seeds (11, 27, 42), and the mean AUC values, standard deviations, and 95% confidence intervals are now reported to reflect performance variability. As the tab transformer was the most efficient model, we conducted a full range of statistical significance tests, including the χ^2 test, the Wilcoxon signed-rank test, the independent t -test, the Mann–Whitney U test, and the DeLong's test. The p -values were <0.05 in all tests, as shown in Table 2, indicating no statistically significant differences between the transformer-based model and the other models.

Although the AUC of 0.97 is considered high, several measures were taken to minimize the risk of overfitting, including patient-level splitting, avoiding data leakage, regular hyperparameter tuning, early termination, and testing on a separate cohort. The slightly lower AUC on

the external dataset, the low variance in seeds, and the well-calibrated probability estimates also contributed to the model's stability and generalizability. Taken together, these analyses enhanced the scientific validity of the results and proved that the model performance is reliable and reproducible.

5.3. Clinical significance

To assess the reliability of the predicted probabilities, we evaluated model calibration using both Brier scores and calibration curves. Among all models, the tab transformer demonstrated the best probabilistic accuracy with a Brier score of 0.0303, followed by SAINT (0.0742) and FT transformer (0.1173), indicating that the tab transformer produced the most well-calibrated risk estimates, as shown in Table 3. The calibration intercepts and slopes further supported this observation, with the tab transformer showing a slope closest to the ideal value of 1, reflecting appropriate scaling of predicted risks. The calibration curves (Figure 12) reveal that the tab transformer tracked the perfect calibration line more closely across most probability bins, particularly in the clinically relevant mid- to high-risk ranges, while SAINT and FT transformers exhibited greater deviation and overconfidence at certain thresholds.

6. Discussion

Our study presents a unified experimental framework integrating transformer-based models, DL, and traditional ML models to enhance the predictive reliability of CVD risk assessment. Specifically, (i) data augmentation significantly improved model generalization by generating balanced datasets through synthetic oversampling, leading to stable training and consistent predictive performance; (ii) comparative evaluation of traditional ML (logistic regression, AdaBoost, Gaussian Naïve Bayes, LDA), DL (GRU, LSTM), and transformer-based models (SAINT, FT, tab transformer) showed that transformer-based models outperformed others in modeling complex biomarker interactions and nonlinear dependencies; (iii) accuracy and loss curves revealed stable convergence, minimal overfitting, and efficient optimization across epochs, confirming strong generalization; (iv) ROC

Table 2. Statistical testing for the tab transformer model

Statistical test	Statistic value	p -value	Interpretation
χ^2 test	0.192	0.423	$p < 0.07$
Wilcoxon signed-rank test	516	0.129	$p < 0.07$
Independent t -test	0.4978	0.628	$p < 0.07$
Mann–Whitney U test	0.628	0.598	$p < 0.07$
DeLong's test	0.437	0.693	$p < 0.07$

Table 3. Brier and calibration score

Model	Brier score	Intercept	Slope
SAINT	0.0742	−1.1823	2.1360
FT transformer	0.1173	−1.7942	1.6233
Tab transformer	0.0303	−1.6547	1.1356

Abbreviations: FT: Feature tokenizer; SAINT: Self-attention and intersample attention transformer.

analysis highlighted the superior predictive accuracy of transformer-based models, with attention mechanisms enhancing class separability, sensitivity, and specificity; (v) fusion experiments combining tab transformer with LSTM and GRU demonstrated that integrating contextual embeddings with temporal learning improved sequential representation and predictive robustness; (vi) grid search optimization identified an optimal configuration—learning rate of 0.00007, batch size of 32, 200 epochs,

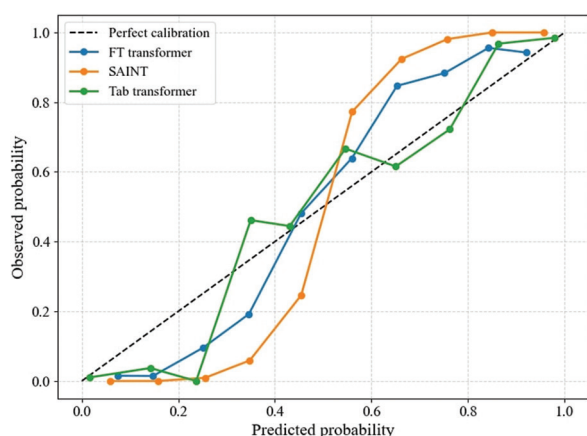


Figure 12. Calibration plots for transformer-based models
Abbreviations: FT: Feature tokenizer; SAINT: Self-attention and intersample attention transformer.

dropout of 0.2, and weight decay of 0.001—ensuring balanced convergence; (vii) multi-fold cross-validation strengthened reliability by minimizing overfitting and improving consistency; and (viii) power analysis using ~2,400 patient samples confirmed sufficient statistical strength, validating that the tab transformer’s performance gains are significant and not due to random variation.

6.1. Benchmarking

The benchmarking in Table 4 summarizes key studies evaluating the performance of artificial intelligence models in CVD risk prediction. It compiles details on dataset sources, patient and feature counts, model types, validation strategies, performance metrics, and power analysis. This structure enabled a clear comparison of traditional ML, DL, and transformer-based approaches across diverse datasets. Maiga and Hungilo²⁸ applied random forest to the Kaggle CVD dataset (70,000 patients, 12 features), achieving 70% accuracy, while Shorewala²⁹ used K-nearest neighbors, random forest, and stacking ensembles, achieving 75% accuracy. Support vector machine-based studies by Unnikrishnan *et al.*,³⁰ Alaa *et al.*,³¹ and Reddy *et al.*³² reported AUC values of 0.71, 0.724, and 0.858, respectively, demonstrating variation across datasets of different scales. DL and attention-based studies have shown improved performance. Bhatt *et al.*³³ achieved 87% accuracy and 0.95 AUC using a multilayer perceptron;

Table 4. Benchmarking table showing studies that implemented machine learning and deep learning models for cardiovascular disease stratification

Citation	Input dataset			Classification model			Performance		Validation
	Source country	Patient number	NOF	Artificial intelligence model	CV	Best Model	Accuracy (%)	AUC [0–1]	Power analysis
28	Kaggle CVD dataset	70,000	12	RF, NB, LR, and KNN	-	RF	70.00	-	No
29	Kaggle CVD dataset	70,000	12	KNN, RF, Stacking	-	Stacking	75.00	-	No
30	-	2,400	9	SVM	K5	SVM	-	0.710	Yes
31	-	423,600	473	SVM, GBM, RF	K10	-	-	0.724	No
32	Cleveland	303	14	SVM	K10	SVM	86.47	0.858	No
33	Kaggle CVD	70,000	14	XGBoost, KNN, MLP	K5	MLP	87.00	0.950	No
34	-	379	97/970	SegNetUNet, SegNet, Unet+SegNet	K5	Unet++	88.90	0.910	No
35	Cleveland	303	13	Self-attention-based transformer	K10	Self-attention-based transformer	96.51	-	No
36	Canada cohort	500	39	RF, RSF	K10	-	-	0.960	Yes
Proposed	Canada cohort	500	39	LR, GNB, LDA, AdaBoost, Tab transformer	K5	Tab transformer	83.76	0.930	Yes

Abbreviations: AUC: Area under the curve; CV: Cross-validation; CVD: Cardiovascular disease; GBM: Gradient boosting machine; GNB: Gaussian Naïve Bayes; KNN: K-nearest neighbors; LDA: Linear discriminant analysis; LR: Logistic regression; MLP: Multilayer perceptron; NB: Naïve Bayes; NOF: Number of features; RF: Random forest; RSF: Random survival forest; SVM: Support vector machine.

Jain *et al.*³⁴ attained 88.9% accuracy and 0.91 AUC using Unet++, and Rahman *et al.*³⁵ reached 96.5% using a self-attention-based transformer. Jamthikar *et al.*³⁶ validated random forest and random survival forest models on the Canada cohort (500 patients, 39 features), reporting 0.96 AUC with power analysis confirmation. The proposed framework integrated ML models (logistic regression, Gaussian Naïve Bayes, LDA, AdaBoost, and XGBoost) with xLSTMmg, achieving 83.76% accuracy and 0.93 AUC on the same cohort using 5-fold cross-validation. The model's temporal learning ability enhanced sequential pattern recognition, offering robust CVD risk prediction and patient stratification.

6.2. A special note on transformer-based models for tabular data

A novel DL model for tackling the challenge of healthcare complexity is a transformer-based model for tabular biomarker data. The attention mechanisms make models such as FT, SAINT, and tab transformer highly efficient in identifying significant interactions among biomarkers and identifying long-term patterns of CVD risk. These models can be used to accurately assess the risk by increasing the sensitivity to small yet clinically significant differences. The tab transformer exhibited high scalability and flexibility and performed better at processing both categorical and continuous data. These models can perform well on noisy, high-dimensional clinical data, suggesting their potential for developing predictive analytics systems in CVD.

6.3. Implication of the study

The clinical implications of this study's findings are significant for CVD risk management. Even small changes in AUC may reduce the number of missed high-risk patients, providing an opportunity to detect subclinical atherosclerosis earlier and to take preventive measures, including statin therapy, lifestyle changes, or follow-up imaging, promptly. The model is improved at capturing silent disease progression using imaging-based biomarkers, such as plaque morphology, which are not considered in conventional risk scores, thereby improving decision-making in preventive cardiology. Better predictive accuracy also reduces false positives, thereby minimizing unnecessary testing and referrals. The combination of these advantages helps improve risk stratification, clinical resource allocation, and clinicians' confidence in excluding or validating high CVD risk.

6.4. Strengths, weaknesses, and extensions

There are several strengths of our solution in CVD risk prediction. (i) Transformer-based models, such as FT, SAINT, and tab transformer, permit the representation

of the nonlinear interactions of biomarkers, allowing the system to learn the complex physiological relationships and obtain a better generalization to different patient phenotypes. (ii) Large-scale hyperparameter optimization and comprehensive power analysis ensure statistically robust and reproducible performance, whereas data augmentation guarantees balanced representation of classes and additional stability of the model. (iii) Imaging-based radiomic features, combined with conventional clinical biomarkers, offer a multimodal view of patient risk that is not provided in conventional clinical scoring systems. (iv) Clinical transparency is improved through the incorporation of interpretability tools, for example, SHAP-based feature importance, which identifies biomarkers that are consistent with established risk factors of CVDs. All these strengths make the suggested framework an effective and clinically relevant cardiovascular risk stratifier.

Despite these strengths, there are several weaknesses. (i) Due to their computational intensity, transformer-based models require a longer time to train, which could make them less applicable in real-time or resource-constrained healthcare settings.^{37,38} This underscores the need for more efficient and streamlined models before clinical deployment. (ii) Our model cannot be directly compared to common CVD risk calculators, such as the Framingham Risk Score or the pooled cohort equations of Atherosclerotic CVD. Our dataset did not provide essential clinical variables required for such comparisons (e.g., total cholesterol, high-density lipoprotein, and detailed treatment history). Consequently, the model demonstrates high predictive performance, but we cannot assess its performance relative to existing clinical instruments used in practice. (iii) Although SMOTE was applied to overcome the issue of class imbalance, the creation of synthetic samples in a relatively small, high-dimensional clinical dataset might not be sufficient to reflect the actual biological variability and might cause slight inflation of model performance. The small size of the natural patient sample also highlights the need for larger, multi-center datasets to confirm the strength and applicability of the proposed strategy.

In the future, the scalability of models and clinical integration should be studied. (i) Future studies should consider compression methods, including model and knowledge distillation, to minimize computation requirements without affecting the model accuracy.^{39,40} Furthermore, multimodal systems that integrate tabular biomarkers with imaging and textual clinical data should be explored to obtain a more efficient and detailed CVD risk assessment.⁴¹

7. Conclusion

Our findings show that transformer-based models, specifically the tab transformer, outperform traditional ML

and DL models. Transformer-based models demonstrated the greatest advantage in the case of heterogeneous biomarkers, such as the combinations of office-based, laboratory-based, and radiomic predictors. The tab transformer's embedding context-aware mechanism enhanced the modeling of nonlinear interactions, particularly among the features GFR, diabetes status, TPA, and medication profiles. These findings underscore that transformer-based models perform better in the multi-source biomarker relationships than either ML or DL models. It was discovered that mechanisms of attention are efficient at processing complex multivariate biomarker data, enabling accurate and reliable prediction of CVD risk. The transformer-based model proposed in the current study is highly scalable and clinically applicable, offering precise risk stratification, which is essential for early diagnosis and intervention. These results suggest that transformer-based models are reliable tools in healthcare analytics, and their future applications are not restricted to CVD but extend to other areas of medical prediction.

Acknowledgments

The data were collected by the Cardiovascular Imaging Network at Queen's University, Canada, and shared with AtheroPoint™, USA, under legal agreements between the two institutes. All software designed and developed in artificial intelligence, DL, is owned and belongs solely to AtheroPoint™. All DL intellectual property is patented and owned by AtheroPoint™, USA.

Funding

None.

Conflict of interest

The authors declare that they have no competing interests.

Author contributions

Conceptualization: Yuvraj Sharma, Neha Gupta, Ekta Tiwari, Jasjit S. Suri

Data curation: Yuvraj Sharma, Esma R. Isenovic, Laura E. Mantella, Amer M Johri

Formal analysis: Narendra N. Khanna, Vandana Kumari, Andrew Nicolaides, Mustafa Al-Maini

Investigation: Yuvraj Sharma, Vijay Rathore, John R Laird

Methodology: Yuvraj Sharma, Neha Gupta

Resources: Ekta Tiwari, Rajesh Singh, Jasjit S. Suri

Supervision: George Kitas, Zoltan Ruzsa, Luca Saba, Jasjit S. Suri

Writing—original draft: Yuvraj Sharma, Gavino Faa, Esma R. Isenovic

Writing—review & editing: Aditya Sharma, Esma R. Isenovic

Ethics approval and consent to participate

Ethics approval for this study was obtained from the Queen's University Health Sciences & Affiliated Teaching Hospitals Research Ethics Board (HSREB), Kingston, Canada. The study was approved under HSREB ID#: 6031758 via a delegated review process.

Consent for publication

All participants provided written informed consent for data collection and publication prior to data collection.

Availability of data

The data used in this manuscript are proprietary to AtheroPoint™ and are not publicly available.

References

1. Aghaei Zarch SM, Dehghan Tezerjani M, Talebi M, Vahidi Mehrjardi MY. Molecular biomarkers in diabetes mellitus (DM). *Med J Islam Repub Iran*. 2020;34:28.
doi: 10.34171/mjiri.34.28
2. Kalantar-Zadeh K, Jafar TH, Nitsch D, Neuen BL, Perkovic V. Chronic kidney disease. *Lancet*. 2021;398(10302):786–802.
doi: 10.1016/S0140-6736(21)00519-5
3. Dai H, Wu S, Huang J, et al. *FT-Transformer: Resilient and Reliable Transformer with End-to-End Fault Tolerant Attention*. [arXiv Preprint]; 2025.
doi: 10.48550/ARXIV.2504.02211
4. Somepalli G, Goldblum M, Schwarzschild A, Bruss CB, Goldstein T. *SAINT: Improved Neural Networks for Tabular Data via Row Attention and Contrastive Pre-Training*. [arXiv Preprint]; 2021.
doi: 10.48550/ARXIV.2106.01342
5. Rahmzadehgervi P, Nguyen HH, Liu R, Mai L, Nguyen AT. *TAB: Transformer Attention Bottlenecks Enable User Intervention and Debugging in Vision-Language Models*. [arXiv Preprint]; 2024.
doi: 10.48550/ARXIV.2412.18675
6. Kim BJ, Nam IW. A review of hybrid LSTM models in smart cities. *Processes*. 2025;13(7):2298.
doi: 10.3390/pr13072298
7. Baibulova M, Aitimov M, Burganova R, et al. A hybrid CNN-GRU-LSTM algorithm with SHAP-based interpretability for EEG-based ADHD diagnosis. *Algorithms*. 2025;18(8):453.
doi: 10.3390/a18080453
8. Zhao S, Wei H, Zhang K. Deep Bidirectional GRU Network for Human Activity Recognition using Wearable Inertial Sensors. In: *2022 3rd International Conference on Electronic Communication and Artificial Intelligence (IWECAI)*. IEEE;

2022. p. 238-242.
doi: 10.1109/IWECAI55315.2022.00054
9. Chen JX, Jiang DM, Zhang YN. A hierarchical bidirectional GRU model with attention for EEG-based emotion classification. *IEEE Access*. 2019;7:118530-118540.
doi: 10.1109/ACCESS.2019.2936817
10. Singh J, Khanna NN, Rout RK, *et al*. GeneAI 3.0: Powerful, novel, generalized hybrid and ensemble deep learning frameworks for miRNA species classification of stationary patterns from nucleotides. *Sci Rep*. 2024;14(1):7154.
doi: 10.1038/s41598-024-56786-9
11. Johri AM, Singh KV, Mantella LE, *et al*. Deep learning artificial intelligence framework for multiclass coronary artery disease prediction using combination of conventional risk factors, carotid ultrasound, and intraplaque neovascularization. *Comput Biol Med*. 2022;150:106018.
doi: 10.1016/j.compbiomed.2022.106018
12. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A simple way to prevent neural networks from overfitting. *J Mach Learn Res*. 2014;15(56):1929-1958.
13. Lin L, Ding L, Fu Z, Zhang L. Machine learning-based models for prediction of the risk of stroke in coronary artery disease patients receiving coronary revascularization. *PLoS One*. 2024;19(2):e0296402.
doi: 10.1371/journal.pone.0296402
14. Teji JS, Jain S, Gupta SK, Suri JS. NeoAI 1.0: Machine learning-based paradigm for prediction of neonatal and infant risk of death. *Comput Biol Med*. 2022;147:105639.
doi: 10.1016/j.compbiomed.2022.105639
15. Sharma Y, Gupta S, Gupta N, *et al*. StockAI 3.0: Ensemble fusion paradigms using novel gating mechanism in long short-term memory architectures for forecasting sentiment-based stock trends. *Soft Comput*. 2025;29(21-22):5803-5829.
doi: 10.1007/s00500-025-10901-8
16. Xu W, He J, Li W, *et al*. Long-short-term-memory-based deep stacked sequence-to-sequence autoencoder for health prediction of industrial workers in closed environments based on wearable devices. *Sensors (Basel)*. 2023;23(18):7874.
doi: 10.3390/s23187874
17. Beck M, Pöppel K, Spanring M, *et al*. *xLSTM: Extended Long Short-Term Memory*. [arXiv Preprint]; 2024.
doi: 10.48550/ARXIV.2405.04517
18. Wilson PWF, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. *Circulation*. 1998;97(18):1837-1847.
doi: 10.1161/01.CIR.97.18.1837
19. Conroy R. Estimation of ten-year risk of fatal cardiovascular disease in Europe: The SCORE project. *Eur Heart J*. 2003;24(11):987-1003.
doi: 10.1016/S0195-668X(03)00114-3
20. Goff DC, Lloyd-Jones DM, Bennett G, *et al*. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: A report of the American college of cardiology/American heart association task force on practice guidelines. *Circulation*. 2014;129(25 Suppl 2):S49-S73.
doi: 10.1161/01.cir.0000437741.48606.98
21. Detrano R, Guerci AD, Carr JJ, *et al*. Coronary calcium as a predictor of coronary events in four racial or ethnic groups. *N Engl J Med*. 2008;358(13):1336-1345.
doi: 10.1056/NEJMoa072100
22. Poplin R, Varadarajan AV, Blumer K, *et al*. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat Biomed Eng*. 2018;2(3):158-164.
doi: 10.1038/s41551-018-0195-0
23. Hossain MM, Ali MS, Ahmed MM, *et al*. Cardiovascular disease identification using a hybrid CNN-LSTM model with explainable AI. *Inform Med Unlocked*. 2023;42:101370.
doi: 10.1016/j.imu.2023.101370
24. Pattanayak S, Singh T. Cardiovascular disease classification based on machine learning algorithms using grid search CV, cross validation and stacked ensemble methods. In: Singh M, Tyagi V, Gupta PK, Flusser J, Ören T, editors. *Advances in Computing and Data Sciences*. Berlin: Springer International Publishing; 2022. p. 219-230.
doi: 10.1007/978-3-031-12638-3_19
25. Konstantonis G, Singh KV, Sfrikakis PP, *et al*. Cardiovascular disease detection using machine learning and carotid/femoral arterial imaging frameworks in rheumatoid arthritis patients. *Rheumatol Int*. 2022;42(2):215-239.
doi: 10.1007/s00296-021-05062-4
26. Blagus R, Lusa L. SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics*. 2013;14(1):106.
doi: 10.1186/1471-2105-14-106
27. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. *JAIR*. 2002;16:321-357.
doi: 10.1613/jair.953
28. Maiga J, Hungilo GG, Pranowo P. Comparison of Machine Learning Models in Prediction of Cardiovascular Disease Using Health Record Data. In: *2019 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*; 2019. p. 45-48.
doi: 10.1109/ICIMCIS48181.2019.8985205
29. Shorewala V. Early detection of coronary heart disease using

- ensemble techniques. *Inform Med Unlocked*. 2021;26:100655.
doi: 10.1016/j.imu.2021.100655
30. Unnikrishnan P, Kumar DK, Poosapadi Arjunan S, Kumar H, Mitchell P, Kawasaki R. Development of health parameter model for risk prediction of CVD using SVM. *Comput Math Methods Med*. 2016;2016:3016245.
doi: 10.1155/2016/3016245
31. Alaa AM, Bolton T, Di Angelantonio E, Rudd JH, Van Der Schaar M. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. *PLoS One*. 2019;14(5):e0213653.
doi: 10.1371/journal.pone.0213653
32. Reddy KVV, Elamvazuthi I, Aziz AA, Paramasivam S, Chua HN, Pranavanand S. Heart disease risk prediction using machine learning classifiers with attribute evaluators. *Appl Sci*. 2021;11(18):8352.
doi: 10.3390/app11188352
33. Bhatt CM, Patel P, Ghetia T, Mazzeo PL. Effective heart disease prediction using machine learning techniques. *Algorithms*. 2023;16(2):88.
doi: 10.3390/a16020088
34. Jain PK, Sharma N, Giannopoulos AA, Saba L, Nicolaides A, Suri JS. Hybrid deep learning segmentation models for atherosclerotic plaque in internal carotid artery B-mode ultrasound. *Comput Biol Med*. 2021;136:104721.
doi: 10.1016/j.combiomed.2021.104721
35. Rahman AU, Alsenani Y, Zafar A, Ullah K, Rabie K, Shongwe T. Enhancing heart disease prediction using a self-attention-based transformer model. *Sci Rep*. 2024;14(1):514.
doi: 10.1038/s41598-024-51184-7
36. Jamthikar AD, Gupta D, Mantella LE, et al. Multiclass machine learning vs. Conventional calculators for stroke/CVD risk assessment using carotid plaque predictors with coronary angiography scores as gold standard: A 500 participants study. *Int J Cardiovasc Imaging*. 2021;37(4):1171-1187.
doi: 10.1007/s10554-020-02099-7
37. Usama M, Ahmad B, Xiao W, Hossain MS, Muhammad G. Self-attention based recurrent convolutional neural network for disease prediction using healthcare data. *Computer Methods Programs Biomed*. 2020;190:105191.
doi: 10.1016/j.cmpb.2019.105191
38. Suri JS, Bhagawati M, Paul S, et al. A powerful paradigm for cardiovascular risk stratification using multiclass, multi-label, and ensemble-based machine learning paradigms: A narrative review. *Diagnostics (Basel)*. 2022;12(3):722.
doi: 10.3390/diagnostics12030722
39. Yeom SK, Seegerer P, Lapuschkin S, et al. Pruning by explaining: A novel criterion for deep neural network pruning. *Pattern Recogn*. 2021;115:107899.
doi: 10.1016/j.patcog.2021.107899
40. Sabih M, Hannig F, Teich J. *Utilizing Explainable AI for Quantization and Pruning of Deep Neural Networks*. [arXiv Preprint]; 2020.
doi: 10.48550/ARXIV.2008.09072
41. Ngiam J, Khosla A, Kim M, Nam J, Lee H, Ng A. Multimodal Deep Learning. In: *Proceedings of the 28th International Conference on Machine Learning, ICML 2011*. 2011:689-696.