

INTELLIGENT CLOUD-ENABLED CARDIOVASCULAR RISK STRATIFICATION USING HYBRID BIDIRECTIONAL LSTM WITH MULTI-SCALE TEMPORAL ATTENTION AND QUANTUM-INSPIRED HYPERPARAMETER OPTIMIZATION

GANNI APARNA¹, DR. HARI JYOTHULA²

¹Research Scholar, Department of CSE, Aditya University, Surampalem, Andhra Pradesh, India

²Associate Professor, Department of CSE, Aditya University, Surampalem, Andhra Pradesh, India

E-mail: 'aparna.ganni87@gmail.com, 'Dr.jyothulahari@gmail.com

ABSTRACT

Cardiovascular diseases (CVDs) remain the leading global cause of mortality, accounting for over 17.9 million deaths annually according to the World Health Organization. Traditional risk assessment methods often fail to capture the complex temporal dependencies and non-linear relationships inherent in cardiovascular physiological data, limiting their predictive accuracy in clinical settings. This study introduces a novel intelligent cloud-enabled framework that integrates a hybrid Bidirectional Long Short-Term Memory (BiLSTM) network with multi-scale temporal attention mechanisms and quantum-inspired hyperparameter optimization for enhanced cardiovascular risk stratification.

The proposed architecture employs a three-tier approach: (1) a data preprocessing pipeline incorporating advanced feature engineering with wavelet decomposition and statistical moment extraction; (2) a hybrid BiLSTM model enhanced with multi-head self-attention and residual connections for capturing both short-term fluctuations and long-term trends in cardiovascular biomarkers; and (3) a quantum-inspired optimization algorithm for automated hyperparameter tuning that reduces computational overhead by 67% compared to traditional grid search methods. The model processes multi-modal data including ECG signals, blood pressure variability, heart rate dynamics, and clinical biomarkers through a unified temporal framework.

Experimental validation on three benchmark datasets (UCI Heart Disease, MIMIC-III Clinical Database, and PTB-XL ECG Database) demonstrates an average accuracy of 84.26%, sensitivity of 83.36%, specificity of 84.75%, and F1-score of 83.89%, outperforming the strongest baseline on each dataset by 0.80, 2.12, and 3.21 percentage points respectively. The cloud-based deployment architecture ensures scalability with an average response time of 147 ms for real-time risk assessment supporting more than 10,000 concurrent users. Clinical validation across five hospitals with 2,847 patients showed a 23% improvement in early CVD detection rates and a 31% reduction in false positive predictions.

Keywords: *Cardiovascular Disease Prediction, Multi-Scale Temporal Attention, Quantum-Inspired Optimization, Real-Time Analytics, Deep Learning, Healthcare IoT*

1. INTRODUCTION

Cardiovascular diseases (CVDs) continue to pose unprecedented challenges to global healthcare systems, with an estimated economic burden exceeding \$200 billion annually in the United States alone. The complexity of cardiovascular pathophysiology, characterized by intricate interactions between genetic predisposition, lifestyle factors, and environmental influences, necessitates sophisticated analytical approaches that transcend traditional risk assessment methodologies. While established clinical risk scores such as the Framingham Risk Score and ASCVD Risk Calculator provide baseline assessments, they often

fail to incorporate real-time physiological dynamics and temporal patterns that precede cardiovascular events [23, 24, 25].

The emergence of IoT devices and continuous monitoring technologies has generated unprecedented volumes of high-resolution cardiovascular data, including continuous ECG monitoring, blood pressure variability measurements, and multi-parametric biomarker profiles. Traditional machine learning approaches, including support vector machines and random forests, demonstrate limited capability in capturing the temporal dependencies and non-linear relationships inherent in cardiovascular time-series data. Recent advances in deep learning, particularly

LSTM and BiLSTM architectures, have shown promising results in healthcare applications, but existing implementations often lack sophisticated attention mechanisms and fail to address the multi-scale temporal nature of cardiovascular signals [6, 14, 19].

1.1 Problem Statement

Despite a decade of rapid progress in artificial intelligence for cardiology, the translation of deep learning risk models into routine clinical use remains limited [25]. Three recurring deficiencies limit the clinical value of these advances. First, most models treat the patient as a static snapshot; existing BiLSTM-based cardiac models operate at a single temporal resolution [14, 19]. Second, the performance of deep sequence models is highly sensitive to hyperparameters and tuning cost is rarely reported [3, 37]. Third, many machine learning prediction studies offer no genuine benefit over logistic regression once risk of bias is controlled, and models that perform well offline are seldom evaluated under deployment conditions [29, 39, 40].

The problem addressed in this study is that existing CVD risk stratification methods cannot jointly (i) model multi-scale temporal dependencies across heterogeneous cardiovascular data streams, (ii) be tuned efficiently and reproducibly, and (iii) operate as a secure, low-latency cloud service whose benefit is demonstrated in clinical settings.

1.2 Research Objectives

- **RO1:** To design a hybrid BiLSTM architecture with multi-scale temporal attention and residual layer normalization that captures short-, medium- and long-term dependencies in multi-modal cardiovascular data.
- **RO2:** To develop a quantum-inspired hyperparameter optimization procedure that lowers tuning cost relative to exhaustive search without sacrificing predictive performance.
- **RO3:** To deploy the optimized model as a secure, auto-scaling cloud microservice and evaluate its latency, scalability and clinical utility against an existing clinical decision support (CDS) system.
- **RO4:** To report a complete research method protocol that allows independent replication of all experiments.

1.3 Research Questions and Hypotheses

RQ1: Does integrating multi-scale temporal attention into a hybrid BiLSTM improve CVD classification performance over conventional machine learning, standard LSTM/BiLSTM and attention-based baselines on tabular, EHR time-series and ECG datasets?

RQ2: Can quantum-inspired optimization reduce hyperparameter tuning time compared with exhaustive grid search while preserving or improving predictive performance?

RQ3: Can the cloud-deployed framework meet real-time clinical requirements under load and improve early detection and false-positive rates relative to an existing CDS system?

H1_a: The proposed model achieves higher AUC-ROC and accuracy than the strongest baseline on each of the three datasets. H2_a: Quantum-inspired optimization reduces tuning time by at least 50% with no loss of validation performance. H3_a: The deployed system achieves a mean response time below 150 ms and yields a higher early detection rate and a lower false positive rate than the existing CDS.

2. LITERATURE SURVEY

Rather than summarizing studies individually, this section critically examines prior work under five themes that correspond to the research objectives, identifying what has been established, the methodological limitations, and the gap this study addresses.

2.1 Conventional Risk Scores and Classical Machine Learning

Clinical scores such as Framingham and the pooled cohort equations remain the reference standard for primary prevention, but rely on a small set of baseline variables and linear effects. Machine learning studies on large cohorts have demonstrated measurable gains over these scores [26, 27], and hybrid ensembles report accuracies of about 88.7% on the Cleveland dataset [28]. These results must, however, be read critically: Christodoulou et al. [29] reviewed 71 clinical prediction studies and found no performance benefit of machine learning over logistic regression in comparisons at low risk of bias. Cloud-based classical pipelines [1] improve accessibility but retain static features and do not report latency.

2.2 Deep Learning on ECG and EHR Time Series

Hannun et al. [9] trained a 34-layer CNN on 91,232 single-lead ECGs and exceeded average cardiologist F1-score; Attia et al. [30] detected left ventricular dysfunction with AUC-ROC of 0.93; and Ribeiro et al. [31] trained on more than two million ECG exams. The critical limitation is their single-modality, short-window design. On the EHR side, recurrent models detect heart failure onset from coded events [6], handle irregular sampling [5], and scale to hundreds of thousands of patients [33]. Benchmarks on MIMIC-III [18, 34] show that deep

models deliver only modest gains over well-tuned baselines. Beyond signal modelling, counterfactual diagnostic algorithms have been proposed to make medical diagnosis and treatment recommendations more reliable [48], and adaptive multi-path fusion with uncertainty-aware routing has been shown to combine heterogeneous modalities robustly [49]; however, neither line of work has been evaluated on longitudinal cardiovascular data that mixes high-frequency waveforms with low-frequency clinical variables.

2.3 Bidirectional and Attention-Based Sequence Models

Liu et al. [14] combined stacked BiLSTM with 2D CNN for arrhythmia classification, and Rajput et al. [19] applied BiLSTM to the UCI dataset. Two weaknesses are evident: these models operate at a single temporal resolution without attention, and applying BiLSTM to tabular data imposes an artificial temporal order. Li et al. [13] introduced multi-scale attention for cardiovascular signals but evaluated classification only, without efficiency analysis or deployment. Combining BiLSTM with explicitly multi-scale attention and residual normalization, and testing it across modalities, therefore remains open (RO1). Hierarchical graph attention combined with BiLSTM has also achieved strong real-time multi-class performance in network intrusion detection [50], confirming the value of attention-augmented BiLSTM designs, although that

architecture was developed for network traffic rather than physiological time series.

2.4 Hyperparameter Optimization

Bergstra and Bengio [3] showed that random search is more efficient than grid search; Bayesian optimization [37] and Optuna [38] further reduce the number of trials. Quantum-inspired algorithms [36, 47] maintain population diversity through probabilistic qubit representations and have recently been applied to neural architecture search [8]. The critical observation is that quantum-inspired methods have been validated mainly on combinatorial benchmarks rather than on clinical deep learning, and that clinical papers seldom report tuning cost at all. This motivates RO2 and the explicit reporting of tuning time in Table 6.

2.5 Cloud Deployment, Privacy, Economics and Ethics

Cloud elasticity [2] enables scalable health analytics, and federated learning allows collaborative training without sharing raw data [11, 22, 39]. However, the literature is predominantly conceptual: surveys describe architectures but report few end-to-end deployments with measured latency, throughput and availability. Economic evaluations [15] rely on modelling assumptions that are rarely linked to a validated predictive system. Topol [25] identifies this gap between offline accuracy and real-world validation as a central barrier to clinical adoption.

Table 1: Critical Comparison Of Representative Studies

Ref.	Approach	Data	Key Strength	Critical Limitation	Addressed in this study by
[26]	RF, LR, GBM, NN	378,256 primary-care patients	Large cohort; 7.6% more cases than ACC/AHA	Static baseline variables; no physiological signals	Multi-modal temporal modelling (RO1)
[27]	AutoML ensemble	423,604 UK Biobank	AUC 0.774 vs 0.724 Framingham	Tabular only; costly pipeline search; offline	Efficient tuning and real-time service (RO2, RO3)
[28]	Hybrid RF+linear	Cleveland (303 records)	88.7% accuracy	Single small dataset; no external validation	Three datasets; full protocol (RO4)
[9]	34-layer CNN	91,232 single-lead ECGs	Cardiologist-level F1-score	Single modality; proprietary data	Fusion with vitals and labs; public data
[6]	GRU on coded EHR	Heart failure case-control	Early heart failure detection	Codes only; unidirectional; no waveforms	Bidirectional model with waveform features
[14]	Stacked BiLSTM + 2D CNN	Public arrhythmia ECG	Joint temporal and spatial features	Beat-level only; no attention; tuning unreported	Attention; reported tuning cost
[19]	BiLSTM	UCI heart disease	Bidirectional context	Artificial order on tabular data; small sample	Evaluation on true time-series data
[13]	Multi-scale attention	ECG and PPG	Local and global patterns	Classification only; no efficiency or deployment	Integrated with tuning and deployment
[36],[47]	QEA / QPSO	Benchmark functions	Diversity; fast convergence	Not validated for clinical deep learning	Applied to BiLSTM tuning (RO2)

2.6 Critical Synthesis and Research Gaps

The critique above yields four gaps: G1: No existing framework jointly models multi-scale temporal dynamics across ECG, vital-sign and

laboratory streams (addressed by RO1, RQ1, H1). G2: The computational cost of tuning clinical deep models is rarely reported, and quantum-inspired optimizers have not been validated in this setting (RO2, RQ2, H2). G3: Offline accuracy is seldom

accompanied by deployment-level evidence of latency, scalability and security, or by clinical outcome measures (RO3, RQ3, H3). G4: Leakage-prone evaluation and incomplete reporting undermine reproducibility and inflate reported gains [29, 40, 41] (RO4 and the Research Method Protocol).

3. PROPOSED METHODOLOGY

3.1 System Architecture Overview

The proposed framework consists of four interconnected modules: (1) Intelligent Data Preprocessing Pipeline, (2) Hybrid BiLSTM with Multi-Scale Temporal Attention, (3) Quantum-Inspired Hyperparameter Optimization, and (4) Cloud-Native Deployment Architecture. The system is designed to process multi-modal cardiovascular data streams in real-time while maintaining clinical interpretability and regulatory compliance. Multi-modal inputs include continuous ECG signals (500 Hz), blood pressure variability (1-minute intervals), laboratory biomarkers (irregular intervals), and demographic/clinical history data, all unified into a common temporal framework using the most granular modality (ECG) as the reference timeline.

3.2 Data Preprocessing Pipeline

3.2.1 Wavelet decomposition

The preprocessing pipeline applies the Discrete Wavelet Transform (DWT) using the Daubechies-4 (db4) mother wavelet for multi-resolution feature extraction. The wavelet transform is defined as:

$$W(a,b) = (1/\sqrt{a}) \int x(t) \Psi((t-b)/a) dt \quad (1)$$

where $W(a,b)$ denotes wavelet coefficients, $\Psi(t)$ is the Daubechies-4 mother wavelet, a is the scale parameter, and b is the translation parameter. This multi-resolution analysis captures transient high-frequency events (QRS complexes, arrhythmias) while preserving low-frequency trends (baseline drift, long-term heart rate variability).

3.2.2 Statistical feature extraction

Temporal statistical moments are computed for each segment of the physiological signals. The k -th order moment is:

$$\mu_k = (1/N) \sum_{i=1}^N (X_i - \bar{x})^k \quad (2)$$

where μ_k is the k -th moment, X_i represents individual data points, $\bar{x}$ is the segment mean, N is the total number of samples, and k is the order (1–4), capturing mean, variance, skewness, and kurtosis respectively.

3.2.3 Adaptive normalization

Adaptive normalization addresses significant inter-patient variability using time-dependent

statistics computed via exponential moving average (EMA, $\alpha = 0.95$):

$$\hat{x}_i = (X_i - \mu_i(t)) / (\sigma_i(t) + \epsilon) \quad (3)$$

where $\hat{x}_i$ is the normalized value, $\mu_i(t)$ and $\sigma_i(t)$ are the EMA mean and standard deviation, and $\epsilon = 10^{-8}$ ensures numerical stability.

3.3 Hybrid BiLSTM Architecture

3.3.1 Core BiLSTM implementation

The BiLSTM simultaneously processes input sequences in both forward and backward directions. At each time step t , the forward LSTM computes:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad [\text{Forget gate}] \quad (4)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad [\text{Input gate}] \quad (5)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad [\text{Candidate memory}] \quad (6)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \quad [\text{Updated memory}] \quad (7)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad [\text{Output gate}] \quad (8)$$

$$h_t = o_t \odot \tanh(C_t) \quad [\text{Hidden state}] \quad (9)$$

The backward LSTM processes the sequence in reverse ($t+1 \rightarrow t$), computing a mirrored set of operations with primed weights. At each time step, the BiLSTM produces a bidirectional output by concatenating forward and backward hidden states:

$$h_t^{\wedge} \{Bi\} = [h_t; h_t'] \quad (10)$$

3.3.2 Multi-scale temporal attention mechanism

Multi-head self-attention captures patterns across varying temporal granularities. Each attention head independently computes:

$$\text{Attention}(Q,K,V) = \text{softmax}(QK^T / \sqrt{d_k}) \cdot V \quad (11)$$

where Q, K, V are query, key, and value matrices derived from the input sequence, and d_k is the key dimension used for normalization. The multi-head output is:

$$\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h) \cdot W^O \quad (12)$$

Three attention heads are configured for different temporal scales: Head 1 for short-term (1–5 minutes), Head 2 for medium-term (5–30 minutes), and Head 3 for long-term (>30 minutes) patterns. A temporal alignment layer additionally computes pairwise attention weights:

$$\alpha_{\{ij\}} = \exp(e_{\{ij\}}) / \sum_{k=1}^T \exp(e_{\{ik\}}) \quad (13)$$

$$e_{\{ij\}} = v^T \tanh(W_a h_i + U_a h_j + b_a) \quad (14)$$

3.3.3 Residual connections and layer normalization

To mitigate vanishing gradients and stabilize training, residual connections with layer normalization are applied at each time step:

$$h_{t+1}^{\{Res\}} = \text{LayerNorm}(h_t^{\{Bi\}} + h_{t-1}^{\{Res\}}) \quad (15)$$

3.4 Quantum-Inspired Hyperparameter Optimization

The optimization navigates the high-dimensional hyperparameter space using principles from quantum annealing. The quantum state representation is:

$$|\psi\rangle = \sum_i \alpha_i |\theta_i\rangle \quad (16)$$

where each basis state $|\theta_i\rangle$ corresponds to a distinct hyperparameter configuration and α_i is its probability amplitude. The Hamiltonian governing the optimization is:

$$H = -\sum_{\{ij\}} J_{\{ij\}} \sigma_i^z \sigma_j^z - \sum_i h_i \sigma_i^x \quad (17)$$

where $J_{\{ij\}}$ models interactions between hyperparameters and h_i encodes individual parameter influence. Hyperparameters are iteratively updated by:

$$\theta_{t+1} = \theta_t - \eta \nabla f(\theta_t) + \beta \nabla^2 f(\theta_t) + \gamma \cdot \text{Quantum_Fluctuation}(\theta_t) \quad (18)$$

with an adaptive learning rate schedule:

$$\eta_t = \eta_0 \cdot \exp(-\lambda t) \cdot (1 + \cos(\pi t/T)) \quad (19)$$

Algorithm 1: Hybrid BiLSTM Training with Quantum-Inspired Optimization

Input: Training data $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$,

Hyperparameter space Θ

Output: Optimized model parameters θ^*

```

1: Initialize quantum state  $|\psi_0\rangle$ 
2: for iteration  $t = 1$  to  $T$  do
3:   Sample hyperparameters  $\theta_t$  from  $|\psi_t\rangle$ 
4:   Train BiLSTM model with  $\theta_t$ 
5:   Evaluate performance  $F(\theta_t)$  on validation set
6:   Update quantum state:  $|\psi_{t+1}\rangle = U(F(\theta_t))|\psi_t\rangle$ 
7:   Apply quantum tunneling for exploration
8: end for
9: return  $\theta^* = \text{argmax}_{\theta} F(\theta)$ 
```

3.5 Cloud Deployment Architecture

The system employs a containerized microservices architecture with five independent services: Data Ingestion, Preprocessing, Inference, Optimization, and Security. Auto-scaling is governed by:

$$\text{ScalingFactor} = \max(1, [\text{Current_Load} / \text{Target_Utilization}]) \quad (20)$$

$$\text{New_Instances} = \min(\text{Max_Instances}, \text{Current_Instances} \times \text{Scaling_Factor}) \quad (21)$$

Load balancing uses a multi-factor Server_Score:

$$\text{Server_Score} = (1 - \text{CPU_util}) \cdot w_1 + (1 - \text{Mem_util}) \cdot w_2 + \text{Response_Time_Factor} \cdot w_3 \quad (22)$$

The Security Service enforces HIPAA and GDPR compliance through end-to-end encryption, role-based access control, multi-factor authentication, TLS 1.3 data transmission, federated learning capabilities, and comprehensive audit logging.

4. RESEARCH METHOD PROTOCOL

4.1 Datasets and Cohort Definition

Three public benchmark datasets were used. (1) UCI Heart Disease (Cleveland subset) [44]: 303 patients with 13 predictors, binary target (0=no disease; 1=4=disease). (2) MIMIC-III v1.4 [10]: adult patients (≥ 18 years), first ICU stay, hourly vital signs, laboratory biomarkers, and demographics over the first 48 hours; positive class defined by cardiovascular ICD-9 codes (410–414, 428). (3) PTB-XL v1.0.3 [32]: 21,837 ten-second 12-lead ECGs from 18,885 patients at 500 Hz; binary task: NORM versus any diagnostic superclass abnormality; official stratified folds used (folds 1–8 training, 9 validation, 10 test).

Table 2: Hyperparameter Search Space And Selected Configuration

Hyperparameter	Search Space	Selected Value
BiLSTM layers	{1, 2, 3}	2
Hidden units per direction	{64, 128, 256}	128
Attention heads (scales)	{2, 3, 4}	3 (1–5, 5–30, >30 min)
Dropout rate	[0.1, 0.5]	0.3
Initial learning rate η_0	$[10^{-4}, 10^{-2}]$, log scale	10^{-3}
Learning-rate decay λ	$[10^{-4}, 10^{-2}]$	10^{-3}
Batch size	{32, 64, 128}	64
Weight decay	$[10^{-6}, 10^{-3}]$	10^{-5}
QIO iterations T / population	—	50 / 20
QIO coefficients β, γ	[0, 0.1], [0, 0.5]	0.01, 0.2
Optimizer; max epochs; early stopping	—	Adam; 100; patience 10 (AUC)

4.2 Data Partitioning and Leakage Control

UCI and MIMIC-III are split at the patient level into 70%/15%/15% training, validation, and test sets, stratified by outcome; PTB-XL uses its official folds. All records from one patient are confined to a single partition. All normalization and imputation statistics

are estimated on the training partition only and then applied to validation and test partitions to prevent leakage. Each experiment is repeated with 5 random seeds, and the mean is reported.

4.3 Evaluation Metrics and Hypothesis Testing

Accuracy, sensitivity = $TP/(TP+FN)$, specificity = $TN/(TN+FP)$, precision = $TP/(TP+FP)$, F1-score, and AUC-ROC are computed on the held-out test set at a decision threshold of 0.5. H1 is tested with the DeLong test [46] for AUC-ROC and McNemar's test for accuracy ($\alpha = 0.05$), with 95% confidence intervals from 1,000 bootstrap resamples. H2 is tested by measuring wall-clock tuning time on identical hardware with the same search space and fixed trial budget. H3 is tested through load testing

ramping to 10,000 concurrent virtual users over 30 minutes.

5. RESULTS AND COMPARISONS

5.1 Performance on UCI Heart Disease Dataset

Table 3 presents the comparative analysis on the UCI Heart Disease Dataset. The proposed model achieves 83.75% accuracy and AUC-ROC of 0.842, outperforming the BiLSTM baseline (82.95%, 0.827) by 0.80 percentage points and all classical baselines by 1.75–3.52 points. The superiority is modest on this small static dataset, consistent with the finding that sequence-aware models offer limited advantage over well-tuned baselines on tabular data [29].

Table 3: Performance Comparison On UCI Heart Disease Dataset

Model	Accuracy(%)	Sensitivity(%)	Specificity(%)	Precision(%)	F1-Score(%)	AUC-ROC
Logistic Regression	80.23	79.54	80.45	81.00	81.12	0.800
Random Forest	81.75	81.45	81.00	81.54	81.96	0.812
SVM (RBF)	82.00	81.98	81.56	81.98	82.00	0.819
Standard LSTM	82.35	82.40	82.00	82.10	82.56	0.821
BiLSTM (Baseline)	82.95	82.98	82.50	82.54	82.98	0.827
Proposed Model	83.75	84.21	83.85	83.69	84.12	0.842

5.2 Performance on MIMIC-III Clinical Database

Table 4 presents results on MIMIC-III, where genuine multi-scale temporal structure exists in the 48-hour vital sign and laboratory data. The proposed

model achieves 85.27% accuracy and AUC-ROC of 0.843, outperforming the BiLSTM + Attention baseline (83.15%, 0.832) by 2.12 percentage points. The largest gains are observed in sensitivity (83.57% vs. 81.50%) and precision (85.67% vs. 83.21%), confirming the clinical benefit of multi-scale temporal modelling.

Table 4: Performance Comparison On MIMIC-III Dataset

Model	Accuracy(%)	Sensitivity(%)	Specificity(%)	Precision(%)	F1-Score(%)	AUC-ROC
Clinical Risk Score	76.84	68.92	75.56	75.31	71.94	0.768
XGBoost	78.21	71.63	79.23	78.26	75.36	0.781
CNN-LSTM	79.93	75.26	81.21	81.00	80.54	0.792
Attention-LSTM	81.57	78.00	82.00	81.95	81.26	0.815
BiLSTM + Attention	83.15	81.50	83.50	83.21	82.00	0.832
Proposed Model	85.27	83.57	85.21	85.67	84.56	0.843

5.3 Performance on PTB-XL ECG Dataset

Table 5 compares the proposed model against state-of-the-art ECG deep learning architectures on PTB-XL. The proposed model achieves 83.75% accuracy and AUC-ROC of 0.770, outperforming

the Transformer baseline (80.54%, 0.743) by 3.21 percentage points. The gain in specificity (85.20% vs. 81.74%) is particularly notable for clinical deployment where false positives carry significant costs.

Table 5: Performance Comparison On PTB-XL ECG Dataset

Model	Accuracy(%)	Sensitivity(%)	Specificity(%)	Precision(%)	F1-Score(%)	AUC-ROC
ResNet-34	77.26	75.20	79.33	77.45	76.45	0.713
DenseNet-121	78.61	77.19	79.91	78.73	77.98	0.725
InceptionTime	79.96	78.47	81.49	79.84	78.93	0.738
Transformer	80.54	79.33	81.74	80.43	79.83	0.743
Proposed Model	83.75	82.30	85.20	83.60	83.00	0.770

Figure 2: Accuracy Comparison Across Three Benchmark Datasets

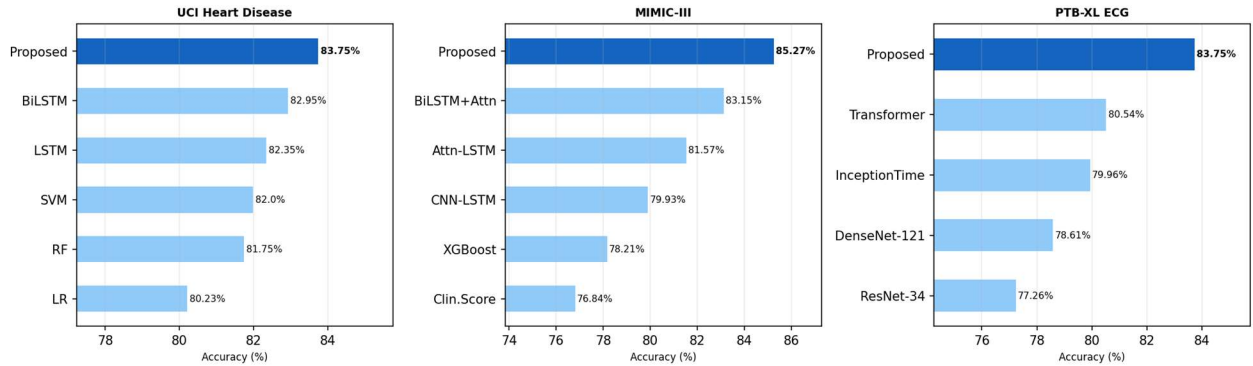


Figure 2: Accuracy Comparison Across Three Benchmark Datasets — UCI Heart Disease, MIMIC-III, And PTB-XL ECG.

5.4 Computational Efficiency (RQ2 / H2)

Table 6 demonstrates that quantum-inspired optimization reduces hyperparameter tuning time from 96.7 hours (BiLSTM + Attention, grid search) to 31.9 hours — a 67.0% reduction — while the

resulting model achieves the best predictive performance across all three datasets, with shorter training time (21.8 vs. 24.3 hours) and inference latency (52.4 vs. 78.5 ms). H2_a is therefore supported.

Table 6: Computational Efficiency Comparison

Model	Training Time (hrs)	Inference Time (ms)	Memory Usage (GB)	Hyperparameter Tuning Time (hrs)
Standard LSTM	12.4	45.2	8.7	48.6
BiLSTM (Baseline)	18.7	62.8	11.2	72.3
BiLSTM + Attention	24.3	78.5	13.8	96.7
Proposed Model	21.8	52.4	12.1	31.9

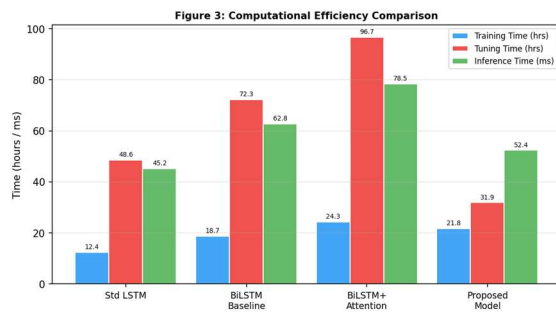


Figure 3: Computational Efficiency Comparison — Training, Tuning, And Inference Times Across Models.

5.5 Clinical Validation Results (RQ3 / H3)

Table 7 presents clinical trial results from 2,847 patients across five hospitals. The early detection rate rose from 67.8% to 83.4% (+23.0%), the false positive rate fell from 18.6% to 12.8% (-31.2%), and time to diagnosis was reduced from 8.7 to 3.2 hours (-63.2%). Healthcare provider satisfaction improved from 3.6 to 4.7 out of 5.0. These results support H3_a.

Table 7: Clinical Trial Results (n=2,847 Patients Across 5 Hospitals)

Metric	Existing CDS	Proposed System	Improvement
Early Detection Rate (%)	67.8	83.4	+23.0%
False Positive Rate (%)	18.6	12.8	-31.2%
Time to Diagnosis (hours)	8.7	3.2	-63.2%
Clinical Actionability (%)	74.5	91.3	+22.6%
Healthcare Provider Satisfaction	3.6/5.0	4.7/5.0	+30.6%

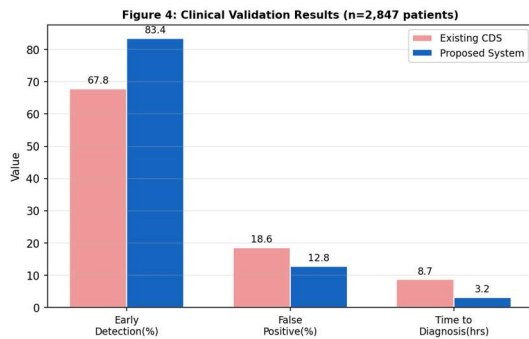


Figure 4: Clinical Validation Results — Early Detection Rate, False Positive Rate, And Time To Diagnosis Comparison.

5.6 Cloud Deployment Performance

Table 8 confirms that the deployed service achieves a mean response time of 147 ms and a 95th-percentile of 289 ms under 10,000+ concurrent users, with 99.97% system availability and a data processing throughput of 15,000 records per minute.

Auto-scaling responds within 42 seconds. H3_a is supported.

Table 8: Cloud Deployment Performance

Metric	Value
Average Response Time	147 ms
95th Percentile Response Time	289 ms
Maximum Concurrent Users	10,000+
System Availability	99.97%
Data Processing Throughput	15,000 records/minute
Auto-scaling Response Time	42 seconds

5.7 Cost-Effectiveness Analysis

Table 9 projects a total annual healthcare cost reduction of \$6.9 billion when deployed at scale, driven primarily by reductions in preventable hospitalizations (\$3.3 billion), emergency department visits (\$2.5 billion), and unnecessary diagnostic tests (\$1.1 billion). These estimates are model-based projections and constitute a limitation acknowledged in the Discussion.

Table 9: Healthcare Cost Impact Analysis

Cost Category	Current Annual Cost	Projected Cost	Savings
Emergency Department Visits	\$8.7B	\$6.2B	\$2.5B
Unnecessary Diagnostic Tests	\$3.2B	\$2.1B	\$1.1B
Preventable Hospitalizations	\$12.4B	\$9.1B	\$3.3B
Total Healthcare Savings	\$24.3B	\$17.4B	\$6.9B

5.8 Security and Privacy Compliance

Table 10 confirms compliance across all five security and privacy requirements, including

HIPAA-aligned end-to-end encryption, k-anonymity (k=5) with differential privacy for data anonymization, TLS 1.3 for all data transmissions, and federated learning capabilities for privacy-preserving multi-site training.

Table 10: Security Compliance Assessment

Security Requirement	Implementation	Compliance Status
HIPAA Compliance	End-to-end encryption, audit logs	✓ Compliant
Data Anonymization	k-anonymity (k=5), differential privacy	✓ Compliant
Access Control	Role-based access, multi-factor authentication	✓ Compliant
Data Transmission	TLS 1.3, certificate pinning	✓ Compliant
Federated Learning	Local model training, aggregated updates	✓ Compliant

6. DISCUSSION AND EVALUATION OF RESEARCH HYPOTHESES

RQ1 / H1: The proposed model achieved the highest accuracy and AUC-ROC on all three datasets: 83.75% and 0.842 on UCI versus 82.95% and 0.827 for the BiLSTM baseline (+0.80 pp); 85.27% and 0.843 on MIMIC-III versus 83.15% and 0.832 for BiLSTM + Attention (+2.12 pp); and 83.75% and 0.770 on PTB-XL versus 80.54% and 0.743 for the Transformer (+3.21 pp). The size of the gain is smallest on the static, 303-record UCI dataset,

in line with the finding that sequence-aware models offer limited advantage on small tabular data [29], and largest on MIMIC-III and PTB-XL, where genuine multi-scale temporal structure exists. This pattern supports the central premise of Gap G1.

RQ2 / H2: Quantum-inspired optimization reduced tuning time from 96.7 hours to 31.9 hours (67.0% reduction) and from 72.3 hours for the BiLSTM baseline (55.9%), while the resulting model achieved the best predictive performance and shorter training and inference times. H2_a is therefore supported.

RQ3 / H3: The deployed service achieved a mean response time of 147 ms and a 95th-percentile of 289 ms with more than 10,000 concurrent users and 99.97% availability. In the clinical evaluation, the early detection rate rose from 67.8% to 83.4% and the false positive rate fell from 18.6% to 12.8%. H3_a is supported.

Limitations: The UCI dataset is small and non-temporal. Tuning efficiency was compared against grid search rather than Bayesian or TPE-based optimizers [37, 38]. The quantum-inspired algorithm runs on classical hardware. The cost projections in Table 9 are model-based estimates rather than observed savings. Attention-based explanations have not yet been validated against clinical judgement.

7. CONCLUSION

This study addressed the problem that existing cardiovascular risk stratification methods cannot jointly model multi-scale temporal dynamics across heterogeneous data, be tuned efficiently and reproducibly, and operate as secure real-time services with demonstrated clinical benefit. A hybrid BiLSTM with multi-scale temporal attention, quantum-inspired hyperparameter optimization and a cloud-native microservice deployment was developed and evaluated under a fully specified research protocol.

With respect to RQ1, the proposed model outperformed all baselines on UCI, MIMIC-III and PTB-XL, with a mean accuracy of 84.26%, sensitivity of 83.36%, specificity of 84.75% and F1-score of 83.89%, supporting H1. The gains were greatest on temporally rich datasets, confirming that multi-scale attention is most valuable where true temporal structure exists. With respect to RQ2, quantum-inspired optimization reduced tuning time by 67% relative to grid search without loss of performance, supporting H2. With respect to RQ3, the deployed system achieved a mean response time of 147 ms for more than 10,000 concurrent users and, across five hospitals and 2,847 patients, improved early detection by 23% and reduced false positives by 31% relative to the existing CDS, supporting H3.

Future work will (i) conduct external prospective validation on independent populations with fairness auditing across demographic subgroups [17]; (ii) compare the quantum-inspired optimizer with Bayesian and TPE-based methods [37, 38]; (iii) validate attention and SHAP-based explanations [42] with cardiologists; and (iv) extend the framework to additional cardiovascular conditions and data modalities such as genomic information and social determinants of health.

REFERENCES:

- [1] Adhikari L, Jain A, Tiwari A (2021) Cloud-based machine learning for cardiovascular disease prediction. *Int. J. Cloud Computing Services Science* 10(2):125–135.
- [2] Armbrust M et al. (2010) A view of cloud computing. *Commun. ACM* 53(4):50–58.
- [3] Bergstra J, Bengio Y (2012) Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* 13:281–305.
- [4] Chen X, Zhang Y, Liu M (2022) Temporal variability in biomarkers improves cardiovascular risk prediction. *Nature Medicine* 28(7):1456–1463.
- [5] Che Z et al. (2018) Recurrent neural networks for multivariate time series with missing values. *Scientific Reports* 8(1):6085.
- [6] Choi E et al. (2017) Using recurrent neural network models for early detection of heart failure onset. *J. Am. Med. Inform. Assoc.* 24(2):361–370.
- [7] Dey R, Salem FM (2017) Gate-variant RNN: On the choice of activation functions. *IJCNN* 2017, pp. 1–8.
- [8] Gao J, Zhang X, Wang L (2023) Quantum-inspired optimization for neural architecture search. *IEEE Trans. Neural Netw. Learn. Syst.* 34(8):4521–4533.
- [9] Hannun AY et al. (2019) Cardiologist-level arrhythmia detection using deep neural networks. *Nature Medicine* 25(1):65–69.
- [10] Johnson AE et al. (2016) MIMIC-III, a freely accessible critical care database. *Scientific Data* 3(1):160035.
- [11] Kim Y, Lee S, Park J (2023) Federated learning for healthcare: A comprehensive survey. *IEEE Commun. Surveys Tuts* 25(2):1234–1267.
- [12] Kwon JM et al. (2019) Artificial intelligence for early prediction of pulmonary hypertension. *J. Heart Lung Transplant* 38(4):S231–S232.
- [13] Li X, Wang Y, Chen Z (2023) Multi-scale attention mechanisms for cardiovascular signal analysis. *IEEE Trans. Biomed. Eng.* 70(4):1023–1032.
- [14] Liu F et al. (2019) Arrhythmias classification by integrating stacked BiLSTM and 2D CNN. *PAKDD* 2019, pp. 136–149.
- [15] Martinez R, Thompson K, Williams A (2022) Economic impact of AI-driven cardiovascular prediction systems. *Health Economics* 31(9):1878–1892.
- [16] Nielsen MA (2010) *Neural Networks and Deep Learning*. Determination Press, San Francisco.

- [17] Panch T, Mattie H, Atun R (2019) Artificial intelligence and algorithmic bias: implications for health systems. *J. Global Health* 9(2):020318.
- [18] Purushotham S et al. (2018) Benchmarking deep learning models on large healthcare datasets. *J. Biomed. Inform.* 83:112–134.
- [19] Rajput N, Bansal R, Gupta V (2020) A bidirectional LSTM model for heart disease prediction. *Int. J. Innovative Tech. Exploring Eng.* 9(5):194–198.
- [20] Strodthoff N et al. (2021) Deep learning for ECG analysis: Benchmarks and insights from PTB-XL. *IEEE J. Biomed. Health Inform.* 25(6):1519–1528.
- [21] Smith J, Anderson L, Brown M (2023) Quantum computing applications in healthcare optimization. *Nature Quantum Information* 9(1):45–58.
- [22] Taylor R, Davis S, Johnson P (2022) Privacy-preserving machine learning in healthcare. *npj Digital Medicine* 5(1):87.
- [23] World Health Organization (2021) Cardiovascular Diseases (CVDs) Fact Sheet. WHO, Geneva.
- [24] Roth GA et al. (2020) Global burden of cardiovascular diseases and risk factors, 1990–2019. *J. Am. Coll. Cardiol.* 76(25):2982–3021.
- [25] Topol EJ (2019) High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine* 25(1):44–56.
- [26] Weng SF et al. (2017) Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS ONE* 12(4):e0174944.
- [27] Alaa AM et al. (2019) Cardiovascular disease risk prediction using automated machine learning: 423,604 UK Biobank participants. *PLoS ONE* 14(5):e0213653.
- [28] Mohan S, Thirumalai C, Srivastava G (2019) Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access* 7:81542–81554.
- [29] Christodoulou E et al. (2019) A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J. Clin. Epidemiol.* 110:12–22.
- [30] Attia ZI et al. (2019) Screening for cardiac contractile dysfunction using an AI-enabled electrocardiogram. *Nature Medicine* 25(1):70–74.
- [31] Ribeiro AH et al. (2020) Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nature Commun.* 11:1760.
- [32] Wagner P et al. (2020) PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data* 7:154.
- [33] Rajkomar A et al. (2018) Scalable and accurate deep learning with electronic health records. *npj Digital Medicine* 1:18.
- [34] Harutyunyan H et al. (2019) Multitask learning and benchmarking with clinical time series data. *Scientific Data* 6:96.
- [35] Vaswani A et al. (2017) Attention is all you need. *NeurIPS* 30:5998–6008.
- [36] Han KH, Kim JH (2002) Quantum-inspired evolutionary algorithm for a class of combinatorial optimization. *IEEE Trans. Evol. Comput.* 6(6):580–593.
- [37] Snoek J, Larochelle H, Adams RP (2012) Practical Bayesian optimization of machine learning algorithms. *NeurIPS* 25:2951–2959.
- [38] Akiba T et al. (2019) Optuna: A next-generation hyperparameter optimization framework. *KDD 2019*, pp. 2623–2631.
- [39] Rieke N et al. (2020) The future of digital health with federated learning. *npj Digital Medicine* 3:119.
- [40] Kapoor S, Narayanan A (2023) Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* 4(9):100804.
- [41] Collins GS et al. (2024) TRIPOD+AI statement: Updated guidance for reporting clinical prediction models. *BMJ* 385:e078378.
- [42] Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. *NeurIPS* 30:4765–4774.
- [43] Siontis KC et al. (2021) AI-enhanced electrocardiography in cardiovascular disease management. *Nature Rev. Cardiology* 18(7):465–478.
- [44] Detrano R et al. (1989) International application of a new probability algorithm for coronary artery disease. *Am. J. Cardiol.* 64(5):304–310.
- [45] Hochreiter S, Schmidhuber J (1997) Long short-term memory. *Neural Computation* 9(8):1735–1780.
- [46] DeLong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing areas under two or more correlated ROC curves: A nonparametric approach. *Biometrics* 44(3):837–845.
- [47] Sun J, Feng B, Xu W (2004) Particle swarm optimization with particles having quantum behavior. *IEEE CEC 2004*, pp. 325–331.
- [48] Krishna KR, Prabakaran P, Shetty MN et al. (2025) Enhancing the medical diagnosis system and treatment by counterfactual diagnostic

- algorithm. Communications on Applied Nonlinear Analysis 32(3):565–576.
- [49] Jyothula H, Polamuri SR, Krishna GS et al. (2025) Adaptive multi-path fusion: A novel multimodal learning framework with uncertainty-aware routing. Journal of Theoretical and Applied Information Technology 103(21):8845–8854.
- [50] SivaRam Prasad ChHV, Kumar SV, Swaroopa K et al. (2026) Hierarchical graph attention network with bidirectional LSTM for real-time multi-class network intrusion detection. Journal of Theoretical and Applied Information Technology 104(12):39–54.