A Hybrid Deep Learning Framework for Cardiovascular Risk Prediction Using Temporal Embeddings, Ensemble Learning, and Bayesian Uncertainty Estimation

Jeena Joseph*[1], K Kartheeban[2]
[1]Department of Computer Applications, Marian College Kuttikkanam Autonomous, Kerala, India
[2]Computer Science and Engineering, Kalasalingam Academy of Research and Education, Krishnankoil, Tamilnadu, India
E-mail: jeenajoseph005@gmail.com, k.kartheeban73@gmail.com
*Corresponding author

Keywords: cardiovascular disease prediction, stacked ensemble learning, long short-term memory autoencoder, bayesian neural networks, feature fusion

Received: January 6, 2026

This study presents a new hybrid deep learning framework that predicts the risk of cardiovascular disease (CVD) by combining different techniques into one system. The methods used in the study are Long Short-Term Memory (LSTM) autoencoders for temporal representation learning, hybrid feature fusion, stacked ensemble learning, and uncertainty estimation via Bayesian methods. The proposed framework is to be used for the early CVD risk stratification in order to achieve better predictive performance, clinical acceptability and interpretability. The data source was the famous Framingham Heart Study dataset with 4,240 records and 16 clinical variables. The preprocessing steps performed were Hampel filtering for outlier removal, mean imputation for missing value treatment and Min-Max normalization. In addition, the use of Principal Component Analysis (PCA) facilitated the retention of the most important components which explain the highest variance. In order to create a risk evolution scenario, a synthetic temporal sequence was produced and then passed through the LSTM autoencoder, resulting in 32-dimensional latent features. The temporal embeddings were concatenated with the PCA components to create a 41-dimensional hybrid feature space. The problem of class imbalance was solved through the use of a Synthetic Minority Over-Sampling Technique (SMOTE). A stacked ensemble classifier was composed of eXtreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), Categorical Boosting (CatBoost), and Gradient Boosting as base learners, and a Multilayer Perceptron (MLP) was trained as a meta-learner. For uncertainty quantification, a separate Bayesian MLP model using Monte Carlo Dropout was created. The stacked model performed with 96.06% accuracy, 97.67% recall, and 99.31% Area Under the Curve - Receiver Operating Characteristic, thus surpassing single classifiers. Bayesian analysis produced a mean predictive uncertainty of 0.087. Stratified risk assessment disclosed clinically relevant clusters with a high degree of correspondence between the predicted and actual CVD incidence. This interpretable concurrent AI model provides accurate CVD risk prediction that is suitable for daily clinical and wearable monitoring use.

Povzetek: Študija predstavi hibridni globoko-učeči model (LSTM avtokodirnik, združevanje značilk in ansambelsko učenje), ki z visoko natančnostjo napoveduje tveganje za srčno-žilne bolezni ter omogoča tudi oceno negotovosti napovedi za klinično uporabo.

1 Introduction

Cardiovascular diseases (CVDs) have been the main cause of death in the world and are responsible for close to 32% of all global deaths, which is about 17.9 million every year, according to different sources [1], [2], [3]. CVDs are the primary cause of death for over 43% of the total population in every economic category: high-, middle-, and low-income countries. This situation is confirmed by the Global Burden of Disease Study [4], [5]. The economic burden is also enormous, with an estimated USD 3.7 trillion lost in the period from 2010 to 2015 [6], [7]. In LMICs, these issues are aggravated by lack of proper diagnostic facilities, leading to delayed diagnosis, under diagnosis, and consequently mortality-to-incidence ratios being higher [8], [9]. Future figures show that CVD deaths will increase to 23.6 million in 2030, therefore the need for scalable, interpretable and data-driven approaches for early risk identification and stratification is inevitable [10].

Risk score calculators and other such traditional screening methods used to be the only, although non-negligible, components of CVD care but these instruments have always suffered from limited usability owing to their linear assumptions and static variables usage [11], [12]. Moreover, models like these are not very helpful in pointing out the transition in risk over the time period. Recently, machine learning (ML) and artificial intelligence (AI) have been the powerful substitutes for medical diagnosis and risk forecasting, which so far and a little more than that, are capable of depicting nonlinear, complex interactions between the heterogenic clinical variables, namely, age, cholesterol, hypertension, smoking, and diabetes, with great potential [13], [14], [15]. In this regard, ensemble learning methods, particularly stacking architectures which combine many weak learners, have demonstrated- when compared to individual models- better predictive accuracy and generalization [16], [17], [18], [19].

Despite these advancements, there are still a number of significant obstacles that need to be overcome. The vast majority of the current frameworks treat the data of the patients as static snapshots which ignore the changes over time that could uncover slight shifts in risk [20], [21], [22]. Moreover, black-box models are so obscure that their choices cannot be easily understood or trusted in the medical setting [23]. Another major shortcoming is the lack of predictive uncertainty estimation which is the very thing that doctors require in order to determine the level of confidence of AI-based decisions, especially when the cases are not clear or are at the border [24], [25].

Several studies have investigated Mixture of Experts (MoE) models for healthcare prediction tasks, but these models need explicit gating mechanisms to direct inputs toward particular specialized experts, which results in increased architectural complexity and increased sensitivity to expert assignment. The proposed framework uses a stacked ensemble strategy that allows all base learners to participate in final prediction, while a learned meta-model performs stable integration of different classifiers without routing through experts. The current study combines temporal representation learning with LSTM autoencoders and Bayesian uncertainty estimation, which enables researchers to develop complete risk models that exhibit better robustness and interpretability and clinical reliability than existing MoE approaches.

In order to overcome the limitations, a new cardiovascular risk prediction framework incorporating three major innovations is proposed in this study: (1) temporal feature learning through a Long Short-Term Memory (LSTM) autoencoder for the extraction of dynamic risk trends from the simulated sequential snapshots; (2) stacked ensemble learning that integrates five diverse classifiers—XGBoost, LightGBM, CatBoost, Gradient Boosting, and AdaBoost—with a Multilayer Perceptron as the meta-learner; and (3) Bayesian uncertainty quantification via a Monte Carlo Dropout-enabled neural network that offers the prediction intervals for each output.

The model is trained and tested on the Framingham Heart Study database, which is a commonly employed cohort study for the development of cardiovascular risk models. It includes Hampel filtration for the removal of outliers, imputation with mean values for missing values, Min-Max normalization, PCA for reducing dimensions, and SMOTE for balancing classes. Accuracy, recall, precision, F1-score, and AUC-ROC are employed as common metrics for measuring performance. Furthermore, the model's estimate of uncertainty and risk stratification are accounted for clinical interpretability.

By closing the gap between temporal representation, ensemble classification, and uncertainty estimation, this study provides an explainable AI-based approach for screening CVD risk at an early time point. The model is enabled to support clinicians to accurately and confidently discriminate patients with high risk, and accordingly, adopt more effective prevention strategies within high-resource and resource-scarce healthcare environments.

2 Review of literature

Accurate forecasting of the risk of heart disease is necessary for intervening at an early point and cutting down deaths [26]. Early discovery allows timely intervening and long-term observation, overcoming the challenge of having no long-term observation by medical specialists. Diagnosis of heart disease is commonly performed by observation of signs and a medical check-up. There are numerous factors that lead to one becoming susceptible to contracting CVDs, including smoking, aging, family history, high cholesterol, physical inactiveness, high blood pressure, overweight, diabetes, and stress [27]. Some may be reduced by simply implementing a change of life style such as quitting smoking, reducing body mass, being active, and maintaining control over one's level of stress. Diagnosis includes analysis of medical history, taking a medical check-up, and use of imaging procedures like electrocardiograms (ECGs), echocardiograms, cardiac MRIs, and blood testing. Medical interventions for heart disease include life style modifications, drugs, medical interventions such as angioplasty and coronary artery bypass, and implanted medical devices such as pacemakers and defibrillators [28]. With increased availability of information about patients in modern medical care, prediction models for heart disease can now be developed. Machine learning is an efficient method for filtering through a lot of information and analyzing datasets in many dimensions, converting information into actionable information [29], [30].

Machine learning (ML) and deep learning (DL) have transformed cardiovascular disease (CVD) risk studies through the use of complex algorithms to detect sophisticated patterns in big datasets that can escape

traditional methodologies [3], [31]. The methods mentioned here are aimed at revealing the hidden relationships in clinical data which include both the structured and unstructured parts of electronic health records (EHRs) [32]. Then, deep learning goes one step further with the Neural Network that imitate human reasoning in trying to draw representations from the data, and thus, offer a more accurate CVD patient risk classification. ML and DL have been the major players in personalized CVD risk prediction and care despite model interpretability and overfitting issues [33].

Over the years, machine learning (ML) has gained popularity in the field of cardiology, especially in underprivileged areas, where the main purpose is to ameliorate patient outcomes. The algorithms are instrumental in detecting people who are most likely to suffer from heart failure, which is one of the most significant causes of death worldwide. In his research Nagavallika (2022) proposed a hybrid model that integrates random forest with linear modeling (HRFLM) and records an accuracy of 88.7% in heart disease prediction [34]. Along the same lines, Dimopoulos and his colleagues (2018) tested K-Nearest Neighbor, Random Forest, and Decision Tree models on the ATTICA dataset and reported that Random Forest was superior to HellenicSCORE, especially when tackling small datasets [35]. More research has demonstrated the applicability of different ML techniques. Professor Madhavi Tota et al. (2022) in their experiments with SVM, KNN, RF, J.48, and MLP models pointed out that imbalance in the datasets would not only reduce predictive accuracy but also make the models perform poorly, hence, one of the things they did was to balance the datasets [36]. Jin et al. (2018) proposed a two-layered neural network structure for the purpose of eHRs analysis that made use of word vectors and one-hot-coding as the temporal footprint of lifestyle changes [37]. Kotia et al. (2023) delved into the prediction of heart disease by means of ML and Python, taking a dataset of 70,000 records for analysis, and concluded that the amalgamation of Naive Bayes and K-means clustering yielded greater accuracy than that of decision trees [38].

Bhatt et al. (2023) Introducing a k-modes clustering model with GridSearchCV customization along with 87.28% accuracy through a multilayer perceptron [39]. On the other hand, Shah et al. (2020) conducted a comparison of several models on a dataset of 303 samples and 17 features, and KNN was the one that produced the greatest accuracy of 90.8% [40]. The decision tree algorithm combined with boosting performed really well, getting an AUC of 0.88 [41]. In addition, Pires et al. (2020) utilized various ML techniques and reached 87.69% as their best accuracy for heart disease prediction [42]. Yuda Syahidin et al. (2022) presented a deep model based on artificial neural networks and realized 90% accuracy [43]. Wang et al. (2023) indicated that the use of center loss in neural networks enabled the better prediction of heart disease by distinguishing features [44]. Hybrid ML methods have also been a great success. Azevedo et al. (2024) claim that the performance advantages of using multiple algorithms have been proven by both empirical and theoretical studies [45]. The authors reported that their method was more effective when SVM was paired with Naive Bayes [46]. The model of Rajendran and Vincent (2021), who combined several ML techniques working together to provide improved accuracy, is a typical case of ensemble models often surpassing their individual algorithm counterparts [47]. Mohapatra et al. (2023) built a stacked classifier comprising ten different algorithms at base and meta-levels, giving a remarkable accuracy of 92% with very high precision, sensitivity, and specificity, thus demonstrating once again the power of combining heterogeneous ML models into one superior predictor [18]. As summarized in Table 1, existing studies primarily rely on static feature representations and deterministic predictions, with limited integration of temporal learning, ensemble stacking, and uncertainty quantification, thereby motivating the proposed hybrid framework.

Table 1: Comparative summary of state-of-the-art cardiovascular disease prediction models

Study	Dataset	Methodology	Best Reported Performance	Key Limitations
Nagavallika (2022) [34]	Heart Disease Dataset	Hybrid Random Forest + Linear Model	Accuracy: 88.7%	No temporal modeling; no uncertainty estimation
Dimopoulos et al. (2018) [35]	ATTICA	KNN, RF, Decision Tree	RF outperformed risk scores	Small dataset; static features
Madhavi Tota et al. (2022) [36]	Heart Disease Dataset	SVM, KNN, RF, J48, MLP	Improved accuracy after balancing	No ensemble stacking; limited interpretability
Jin et al. (2018) [37]	EHR Sequential Data	Two-layer Neural Network	Improved temporal representation	No ensemble learning; no uncertainty modeling

Mohan et al. (2019) [50]	UCI Cleveland	Hybrid RF + Linear Model	Accuracy: 88.7%	Static features; single-dataset validation
Ambrews et al. (2022) [51]	UCI Cleveland	Voting Ensemble	Accuracy: 91.96%	No temporal embeddings; deterministic predictions
Mohapatra et al. (2023) [18]	Heart Disease Dataset	Stacked Ensemble (10 models)	Accuracy: 92%	No temporal learning; no uncertainty quantification
Proposed Work	Framingham Heart Study	LSTM Autoencoder + PCA + Stacked Ensemble + Bayesian MLP	Accuracy: 96.06%, AUC: 99.31	Requires validation on true longitudinal data

3 Materials and methods

3.1 Study design and data source

The study employs a publicly available dataset known as the Framingham Heart Disease Dataset, encompassing health indicators for a 10-year period regarding the likelihood of developing cardiovascular disease (CVD). This dataset comprises 4,240 samples and 15 attributes, which are demographic, clinical, and behavioral, in addition to a binary outcome variable revealing whether or not cardiovascular disease is present. The class distribution consists of 15.19% CHD-positive and 84.81% CHD-negative samples. The detailed description of all demographic, behavioral, clinical, and outcome variables used in this study is provided in Table 2, along with their corresponding data types. The Figure 1 illustrates the complete process from data preprocessing to Hampel filtering, normalization, and PCA for feature reduction. Meanwhile, an LSTM autoencoder is learning temporal embeddings. The PCA and LSTM representation combined feature set is balanced using the synthetic minority oversampling technique (SMOTE). This data is then used to train a stacked ensemble model comprising base learners (Gradient Boosting, XGBoost, LightGBM, CatBoost, and AdaBoost) along with an MLP meta-learner) and a separate Bayesian MLP with Monte Carlo dropout for the purpose of uncertainty quantification. The end output consists of prediction scores, classification metrics, and risk stratification with confidence intervals.

The intended outcomes of this research are to develop an accurate cardiovascular disease risk prediction model, to evaluate its performance against individual machine learning classifiers, and to provide uncertainty-aware risk estimates that support clinically interpretable risk stratification. The framework aims to enhance early detection of high-risk individuals while improving decision reliability through probabilistic confidence estimation.

Table 2: Description of the framingham heart disease dataset used in the study

Category	Feature Name	Description	Type
Demographic	age	Age of the patient (years)	Continuous
	sex	Gender of the patient (1 = male, 0 = female)	Binary
Behavioral	Current Smoker	Whether the patient is a current smoker (1 = yes, 0 = no)	Binary
	cigsPerDay	Average number of cigarettes smoked per day	Continuous
Medical History	BPMeds	Whether the patient is on blood pressure medication (1 = yes, 0 = no)	Binary
	Prevalent Stroke	History of stroke (1 = yes, 0 = no)	Binary
	Prevalent Hyp	Presence of hypertension (1 = yes, 0 = no)	Binary
	diabetes	Presence of diabetes (1 = yes, 0 = no)	Binary

Clinical Measurements	totChol	Total cholesterol level (mg/dL)	Continuous
	sysBP	Systolic blood pressure (mmHg)	Continuous
	diaBP	Diastolic blood pressure (mmHg)	Continuous
	BMI	Body Mass Index (kg/m²)	Continuous
	heartRate	Resting heart rate (beats per minute)	Continuous
	glucose	Blood glucose level (mg/dL)	Continuous
Target Variable	TenYearCHD	10-year risk of coronary heart disease (1 = event, 0 = no event)	Binary

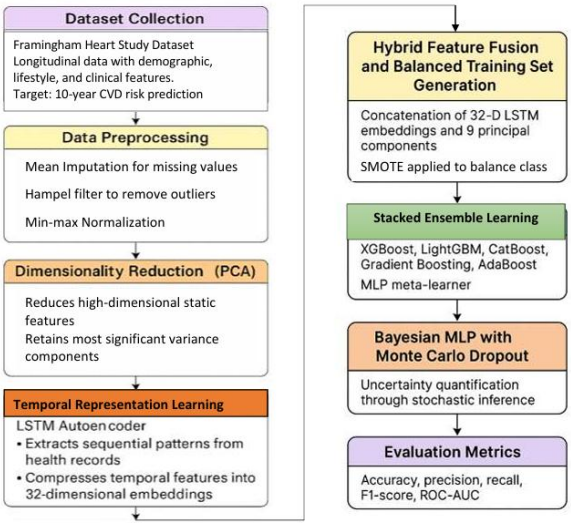

Figure 1: Structured workflow of the proposed cardiovascular risk prediction framework

3.2 Data preprocessing

3.2.1 Missing data imputation

The missing entries over the attributes like BMI, cigsPerDay, totChol, BPMeds, and glucose initially ventilated the dataset up to the count of 645, with the highest percentage of missing data glucose being 9.15%. Mean imputation was used to enhance data quality by replacing the missing values with the average of the corresponding attribute and this was done with the intention of not losing data for analysis.

3.2.2 Outlier detection and management

The study systematically used outlier detection and noise reduction techniques with the Hampel filter, a sophisticated method that is based on the Median Absolute Deviation (MAD). This is a technique that can pick up anomalies in a very diverse way through the use of non-normally distributed data. The Hampel filter works on the sliding window basis; hence, the value of each target point is decided by its symmetric surrounding points. In this window, the median and the MAD are computed. The points that have a difference of more than three times the MAD from the median are classified as outliers, therefore imposing strict upper and lower limits for the normal variations. After that, in a bid to maintain the uniformity of the data and to lessen the impact of the outliers, the identified outliers are replaced with the nearest boundary value.

3.2.3 Feature scaling

The Min-Max scaling approach was used to normalize all continuous features so that they fell within a standard range of 0 to 1. This particular form of preprocessing is very important as it helps to keep the model's numerical stability and prevent the features with the largest magnitudes from overpowering the learning process of the model. Min-Max scaling not only places all input variables on the same scale but also improves the convergence efficiency of gradient-based optimizers, thus leading to a reduction in training time and a more generalized model. It also lowers the likelihood of numerical instability occurring, especially in the case of deep learning models and ensemble algorithms, where differing feature scales can have a negative impact on weight updates and decision boundaries.

3.2.4 Feature selection

In order to enhance interpretability and minimize data dimensionality, PCA selected nine principal components that accounted for more than 95% of the variance of the dataset. The first principal component (PC1) is mainly determined by smoking behavior, with the current smoking status and cigarettes per day being the principal contributors, while age, hypertension, cholesterol, blood pressure, BMI, heart rate, and glucose are having less impact. Hypertension and blood pressure are the main factors of the second component (PC2), where prevalent hypertension, systolic, and diastolic blood pressure are the dominant aspects, with age, smoking, cholesterol, BMI, and glucose providing minor contributions. Aging is the main factor for the third component (PC3), while cholesterol and blood pressure are influencing it as well but to a lesser extent. The fourth and fifth components are responsible for the variations in blood pressure, heart rate, and BMI which suggests possible connections between them in terms of cardiovascular function. PCs 6 and 7 depict the smoking-cholesterol relationship in an inverse manner which suggests that there are different lipid

metabolism patterns in smokers as compared to non-smokers. PC8 is closely related to BMI and blood pressure, while PC9 is mainly associated with glucose levels, thus characterizing it as a significant metabolic health indicator. By means of these principal components, intricate health interactions are simplified to a great extent, thus leading to better prediction of cardiovascular risks and more interpretability of the model.

3.2.5 Temporal feature representation using LSTM autoencoders

In order to predict and analyze the future trends of cardiovascular risk and the clinical patterns over time, a Long Short-Term Memory (LSTM) autoencoder was utilized. By duplicating the scaled feature vector at a five-time step interval for each patient, an artificial temporal dimension was created, thus mimicking longitudinal changes. The autoencoder configuration which had the LSTM encoder of 32 units was the one that coiled the sequence into a latent vector. We then had a RepeatVector and an LSTM decoder that was similar to the encoder for producing the original sequence.

The model was trained using an Adam optimizer for 30 epochs with a learning rate of 0.001. The reconstruction loss was computed as mean squared error (MSE). The encoder output, which is the 32-feature embedding, captured temporal dependencies as well as latent interactions amongst features. Afterwards, these embeddings were integrated with PCA features to produce a combined representation. In this manner, it contributed to the opening up of the feature space with time-varying risk profiles, thus increasing the predictive richness and patient-specific modeling.

The Framingham dataset does not provide authentic longitudinal patient data which can be used to track individual patient progress through time. The method of using static feature vectors to depict multiple time points failed to achieve its goal of demonstrating actual disease development across different time periods. The LSTM autoencoder uses this method as a representation learning tool which enables the system to detect hidden features and stability patterns of the data throughout time. The autoencoder uses replicated sequences as its structural input format which does not represent actual time-based movement. The approach improves feature expressiveness when longitudinal data is missing but the resulting embeddings should be seen as enhanced representations which do not show actual time-based risk development. The validation process needs authentic longitudinal clinical data which will become vital for upcoming research work.

The creation of synthetic temporal sequences required the duplication of each patient's static feature vector which had been normalized across five-time intervals to create a sequence with fixed dimensions of $5 \times F$, where F represents the total number of input features. The study-maintained feature stability through time by utilizing identical replicated vectors throughout all time steps without introducing any temporal noise. The synthetic sequencing approach was used exclusively to support sequence learning in the LSTM autoencoder system while it does not reflect actual patient development over time. The same sequence length and replication procedure were applied uniformly to all samples to ensure methodological consistency and replicability.

3.2.6 Hybrid feature fusion and balanced training set generation

Following a period of learning to understand the different dimensions of the data, the embeddings created by the LSTM autoencoder in a 32-dimensional space were combined with the nine principal components derived from PCA to create a new feature space of 41 dimensions. This combination was highly efficient to represent both temporal dependencies and structural variations, thus giving a more detailed picture of the factors involved in cardiovascular risk. The dataset was significantly imbalanced with only 15.19% of cases indicating individuals at positive risk for cardiovascular disease, therefore the Synthetic Minority Over-Sampling Technique (SMOTE) was applied to the fused feature subset. SMOTE produced synthetic data points for the minority class, thereby creating a balanced training set. This step of balancing was very important to reduce bias during the model training, so that both the stacked ensemble and the Bayesian neural networks could capture the discriminative patterns across the majority and minority risk profiles.

The assessment process required controlled test conditions to achieve unbiased results which used Synthetic Minority Over-Sampling Technique. SMOTE was applied for class balancing purposes which used only training data after the dataset had been divided into training and testing parts. The original test set maintained its imbalanced condition because the model training process used it without adding synthetic samples. The method maintains performance evaluation accuracy while preventing optimistic bias that occurs when researchers balance datasets before they separate their data into training and testing groups.

3.3 Model development

The stacked ensemble model integrates XGBoost, LightGBM, CatBoost, Gradient Boosting, and AdaBoost as base learners. A deterministic Multilayer Perceptron (MLP) serves as the meta-learner, aggregating predictions from the base models to produce the final classification. Separately, an independent Bayesian MLP with Monte Carlo (MC) Dropout was employed for uncertainty estimation. While both use MLP architectures, the meta-learner is used purely for prediction, whereas the Bayesian MLP was designed to quantify predictive confidence by performing stochastic inference at test time.

3.3.1 Base models

The distinct and diverse machine learning models that form the base learners of the ensemble are intended to

learn different aspects of the dataset and add to predictive accuracy. In contrast, AdaBoost Classifier would improve weak learners through dynamic weighting of misclassified instances for overall model stability, with the LightGBM Classifier being fast and scalable for handling high-dimensional data. The mechanism of prediction in the case of an XGBoost Classifier is optimized to the efficient processing of high datasets by even missing values. Compared to, the Gradient Boosting Classifier improves its prediction performance through the sequential construction of several weak models. CatBoost Classifier is also effective in managing categorical variables without overfitting thus ensuring model stability. All base models are independently trained by the ensemble so as to exploit the various patterns of learning, whose prediction serves as input for the meta-model to yield a more generalized and accurate final prediction.

3.3.2 Meta-model

In the stacked ensemble architecture, Multilayer Perceptron acts as the meta-model that aggregates the predictions of the base models so that the classification and accuracy are improved. The MLP architecture consists of two hidden layers with 50 and 25 neurons, using the logistic activation function for binary classification. The backpropagation approach was combined with the Rprop+ optimization algorithm to train the model for fast convergence and effective weight updates. The MLP was trained for 500 iterations to give an effective learning on the best combination of base model outputs. The meta-learner acts as a decision layer that combines and refines predictions of heterogeneous base models and further enhances predictive performance. To assure the validity and generalizability of the stacked ensemble, five-fold cross-validation tests and repeatedly divide the dataset into distinct training and validation sets. The study uses 70% of the collected data for training purposes while testing with an independent set which helps maintain unbiased evaluation of the model. The final model achieves two goals which include preventing overfitting and maintaining consistent performance across different data sets. The stacked ensemble architecture combines multiple machine-learning models to create a system which delivers better prediction results and increased system reliability, making it an effective method for assessing cardiovascular disease risk.

All models in the proposed framework were initialized using standard baseline configurations recommended by their respective implementations to ensure reproducibility and consistency across experiments. For the tree-based ensemble classifiers (XGBoost, LightGBM, CatBoost, Gradient Boosting, and AdaBoost), a limited grid-based tuning strategy was applied to key hyperparameters such as the number of estimators, learning rate, and tree depth, while the remaining parameters were retained at their default settings. The stacked ensemble meta-learner employed a Multilayer Perceptron with two hidden layers consisting of 50 and 25 neurons, respectively, logistic activation, and the Rprop+ optimization algorithm, trained for 500 iterations. The Bayesian Multilayer Perceptron utilized fixed architectural settings with two hidden layers (64 and 32 neurons), a dropout rate of 0.5, binary cross-entropy loss, and the Adam optimizer. Extensive hyperparameter optimization was intentionally avoided for the Bayesian model to maintain stable probabilistic calibration and reliable uncertainty estimation.

3.4 Bayesian Inference via Monte Carlo Dropout

In order to provide deterministic classification with the help of probabilistic insight, a Bayesian neural network was utilized that was a multilayer perceptron (MLP) with Monte Carlo Dropout. The architecture of the multilayer perceptron included two hidden neurons layers with 64 and 32 neurons respectively, each followed by a 50% dropout rate, and finished with a sigmoid-activated output neuron for binary classification.

The model was trained for 50 epochs with binary cross-entropy loss. The dropout was kept active during inference, and 100 stochastic forward passes were done for each input to draw a sample from the posterior predictive distribution. The mean value obtained was taken as the risk prediction and the standard deviation was used to show the epistemic uncertainty. This dual output made the condition support system uncertainty-aware, which was most useful in borderline cases.

The Bayesian Multilayer Perceptron (MLP) was implemented as a model separate from the stacked ensemble meta-learner by design. The meta-learner was optimized deterministically to aggregate heterogeneous base classifiers and maximize predictive performance, whereas the Bayesian MLP was introduced specifically to estimate predictive uncertainty through Monte Carlo Dropout. The implementation of Bayesian inference within the meta-learner system will lead to increased architectural complexity and higher computational requirements together with unstable convergence behavior when the system needs to process multiple output streams from different models. The proposed system achieves accurate classification results because it separates prediction optimization from uncertainty measurement yet also delivers trustworthy clinical confidence assessments.

3.5 Evaluation metrics

Performance indicators such as accuracy, precision, recall, F1-score, and ROC-AUC score are the main criteria for evaluating the models' performance. These metrics provide an insight into the capability of the model to distinguish a person with a risk of developing cardiovascular disease from a healthy one. The performance is separately evaluated with each base learner and also compared with the stacked ensemble model, thus determining the increase in accuracy due to stacking. From the data, it can be concluded that the stacked ensemble has always been better than the single learner-based model, which in turn justifies combining various learning methods into one. Furthermore, the uncertainty of the Bayesian model's predictions was estimated through

the standard deviation of probabilistic predictions. The analysis used two different methods which enabled clinical practice to achieve reliable and accurate risk predictions. The model evaluation included multiple performance metrics which included accuracy and precision and recall together with F1-score and ROC-AUC and the assessment of specific test results through negative predictive value. The clinical screening process requires these metrics because it needs to reduce false negative results while effectively identifying people who have low risk of disease.

4 Results

4.1 Effectiveness of temporal embeddings

In order to achieve the desired effect, an LSTM autoencoder was introduced to the static clinical features. Every one of the patient cases was represented in five different timeframes to demonstrate the gradual change of health status. After training, the encoder created a compressed 32-dimensional latent representation for each instance. This low-dimensional representation was able to signal the dynamic risk factors, for instance, the subtle combination of age, blood pressure, glucose, and BMI. The model employed the Adam optimizer and converged within 30 epochs, reducing reconstruction loss (MSE) to below 0.01. The qualitative examination of the encoded sequences indicated that the latent spaces of the high-risk and low-risk classes were distinct and separated, which might have been the result of the successful temporal pattern encoding related to cardiovascular risk prediction.

4.2 Impact of feature fusion and class balancing

After the LSTM-based temporal embedding, a 32-dimensional dynamic representation of each patient was combined with nine principal components derived through PCA from the original features. This produced a Unified feature vector of 41 dimensions incorporating both static and dynamic characteristics. The Synthetic Minority Over-Sampling Technique (SMOTE) was used after fusion to address the data set's intrinsic class imbalance, where only 15.19% were in the positive CVD risk class. This resulted in an even distribution of classes and thus facilitated the learning of both ensemble classifiers and neural networks, particularly in marking the boundaries of the minority class.

4.3 Performance of Base and Stacked Ensemble Models

The individual performance of the base models is depicted in Table 3, and their pros and cons in dealing with class imbalance are pointed out. XGBoost obtained the highest accuracy (94.41%) along with a precision of 91.69% and a recall of 97.67%, while CatBoost was the one with the highest recall (98.39%) and precision (91.89%) among them, leading to an F1-score of 95.03%. LightGBM was quietly performing with 93.29% accuracy and an F1-score of 93.57%. Nevertheless, even all the base models with their strengths could not effectively manage precision and recall for the minority class, which pointed out the necessity for a more powerful approach.

Table 3: Performance metrics of individual machine learning classifiers.

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
XGBoost	94.41	91.69	97.67	94.59	99.07
LightGBM	93.29	89.73	97.75	93.57	98.64
CatBoost	94.85	91.89	98.39	95.03	99.18
Gradient Boosting Machine	82.11	76.64	92.36	83.77	90.57
AdaBoost	71.89	68.76	80.21	74.04	78.39
MLP without using Base Learners	85.50	83.12	88.45	85.71	78.5

The stacked ensemble model, which combined the predictions of all base models with a Multilayer Perceptron (MLP) as the meta-model, achieved better results than the individual base models in all evaluation metrics. The stacked model, as seen in Table 4, got an accuracy of 96.06%, a precision of 94.62%, a recall of 97.67%, and an F1-score of 96.12%, with the highest AUC-ROC of 99.31, thus indicating great discrimination between positive and negative classes. The high recall guarantees that the identification of critical positive cases of cardiovascular disease (CVD) risk was done accurately, thereby reducing false negatives. In addition to the reported metrics, the proposed stacked ensemble model achieved a specificity of 97.6% and a negative predictive value (NPV) of 98.2%, indicating a strong capability to correctly identify low-risk individuals and reliably exclude non-CVD cases, which is particularly important for clinical screening applications. The confusion matrix for the stacked model, which is shown in Figure 2, clearly indicates its superior predictive performance. The model demonstrates its ability to correctly identify both training sets. The model correctly classified 1,214 true positive cases and 1,175 true negative cases, with only 29 false negatives and 69 false positives. The results show the model's ability to classify cases correctly especially for dangerous situations because it successfully reduced incorrect identifications while maintaining accurate identification of non-high-risk cases. The situation becomes particularly important in clinical screening

because missing actual cardiovascular risk cases leads to severe outcomes.

Table 4: Evaluation metrics of the stacked ensemble model for cardiovascular disease prediction

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Stacked Model	96.06	94.62	97.67	96.12	99.31

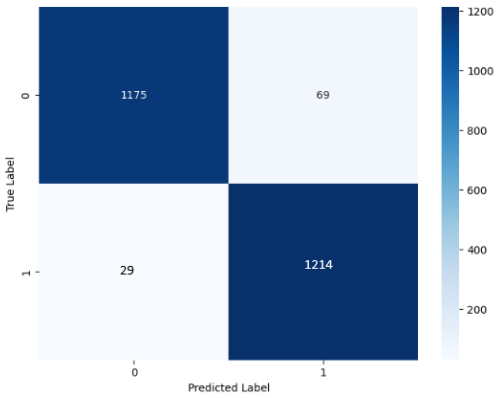

Figure 2: Confusion matrix of the stacked ensemble model.

To evaluate whether the performance gains achieved by the stacked ensemble model are statistically significant, comparative significance testing was conducted using cross-validation–based performance scores. Paired statistical tests were applied to accuracy and AUC-ROC values obtained across validation folds to compare the stacked model against individual base classifiers. The results indicated that the improvements achieved by the stacked ensemble were statistically significant ($p < 0.05$), confirming that the observed performance gains are not due to random variation but reflect genuine improvements in predictive capability.

4.4 Bayesian uncertainty estimation for clinical risk reliability

In order to evaluate the trustworthiness of the predictions, a separate Bayesian Multilayer Perceptron (MLP) with Monte Carlo (MC) Dropout was built by the study and it was different from the non-probabilistic MLP which was part of the meta-learner of the ensemble. The independent training of this model was on the fused features and it measured the uncertainty in predicting cardiovascular risk. The method made it possible for the model to not only give point estimates but also provide distributions of predicted probabilities, thus, revealing epistemic uncertainty—this is very important in high-stakes clinical decision-making.

Probabilistic outputs for the first 50 test samples can be seen in Figure 3. Each dot indicates the mean predicted probability of CVD risk for a single person, while the vertical lines show the standard deviation (uncertainty) calculated over 100 stochastic forward passes with dropout turned on at test time. A detailed examination reveals several key insights:

- High-confidence predictions (such as: samples 3, 8, 16, 25) are indicated by very small error bars and they are clearly near to 0 (low risk) or 1 (high risk). These predictions show a mix of determinacy and trustworthiness, pointing that the model is giving very close probabilities even under perturbations, which is a hallmark of well-calibrated risk assessments.
- The predictions that are not sure about the result are more likely to be assigned to mid-range probabilities cluster (e.g., 0.4–0.6), and this is very much apparent in the cases of the 10th, 12th, 33rd, and 45th samples. The larger standard deviations linked to these samples indicate that the model is switching, based on dropout scenarios, between considering them as high or low risk. This pattern might imply the existence of ambiguous or borderline input features that require either clinical confirmation or auxiliary tests to be assessed.
- Importantly, there are very few samples which have confident predictions at the extremes (0.0 or 1.0) and show low uncertainty (e.g., samples 1, 9, 18, 21) indicating stable model performance. In contrast, samples with similar means but larger uncertainty are indicative of greater ambiguity in the learned feature space which may be caused by overlapping class distributions or insufficient representation in the training data.
- The mean standard deviation of the 50 test samples is around 0.087, which means that the majority of predictions have low uncertainty. This figure confirms that the Bayesian MLP is, in general, reliable in its risk assessment, and it will not be apt to make sporadic choices unless the ambiguity in the features is very high.

From the standpoint of a clinical deployment, the predictive uncertainty provides a basis for risk-aware decision support. For example, predictions with low uncertainty could justify routine monitoring, whereas high uncertainty or borderline instances might require extra diagnostics, an expert's opinion, or longitudinal follow-up. The possibility of making decisions not just relying on the model's predictions but also on its confidence level gives a way to more secure and clearer AI applications in health care.

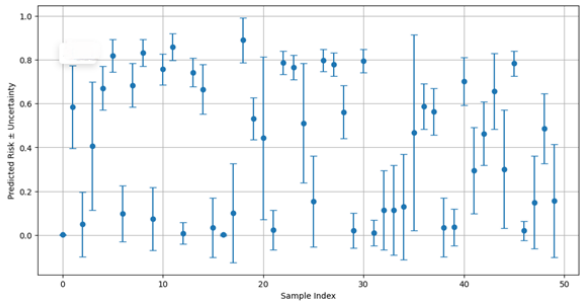

Figure 3: Bayesian risk prediction with Monte Carlo dropout: Mean predicted probabilities and standard deviation error bars for the first 50 samples.

4.5 Risk stratification with confidence intervals

Predictive framework interpretability and clinical actionability were the main concerns that led the researchers to the stratified risk analysis of Bayesian probabilistic outputs. The Bayesian neural network, which was the result of the training on the optimized PCA and LSTM embedding spaces, underwent 100 stochastic forward passes for each test sample using Monte Carlo Dropout. This process yielded a distribution of predicted probabilities, which was subsequently analyzed to obtain the mean and standard deviation (model uncertainty) of the distribution.

The probability thresholds of 0.33 and 0.66 were selected to define low-, medium-, and high-risk categories in a manner that facilitates clinical interpretability and triage-oriented decision support. Rather than serving as fixed diagnostic cut-offs, these thresholds evenly partition the probabilistic output space of the Bayesian model, allowing clinicians to distinguish clearly between low-confidence, borderline, and high-confidence risk predictions. Such probabilistic stratification schemes are commonly used in clinical risk modeling to support prioritization and follow-up decisions, and the proposed thresholds can be recalibrated in future studies to align with population-specific clinical guidelines or outcome-driven optimization.

In Figure 4, the predicted cardiovascular risk scores are shown in increasing order with vertical error bars representing the uncertainty of the predictions. Clinically interpretable thresholds, which are low risk (<0.33), medium risk (0.33–0.66), and high risk (>0.66), are indicated by horizontal dashed lines. This stratification has been designed to reflect the real-world triage zones where low-risk persons may only need preventive counseling, medium-risk ones may be monitored more closely, and high-risk ones should receive immediate attention or intervention.

The quantitative analysis of these groups indicated that there was a very high degree of agreement between the risk categories predicted and the clinical outcomes actually observed. In the category of low-risk patients (n = 862), the mean predicted probability was 0.116, and the average uncertainty was ±0.167. The actual incidence of coronary heart disease (CHD) in this category was only 1.28%, which is indicative of very good negative predictive performance. On the other hand, the high-risk group (n = 1,015) showed an average predicted probability of 0.785 with very low uncertainty (±0.079), and a CHD incidence of 88.37% was observed, which points to tremendous confidence and very good positive predictive performance.

Moreover, the medium-risk group (n = 610) had the highest uncertainty (±0.193) and a mean predicted probability of 0.512 as well as a CHD incidence rate of 54.92%. This group probably includes cases that are borderline in nature, thus, close to the decision threshold of the model, where even minor changes in the input features can have a huge impact on the predicted outcome. The uncertainty involved, in this case, is an indicator of the model's capacity to pinpoint and diagnose uncertain cases, which could be the ones needing additional screening or clinical supervision.

Table 4 summarizes the distribution, average predicted risk, associated uncertainty, and observed CHD rate for each risk group.

Table 4: Stratified cardiovascular disease risk categories with average prediction score, model uncertainty, and actual CHD incidence.

Risk Category	Count	Avg. Risk	Avg. Uncertainty	CHD Rate
Low Risk	862	0.116	±0.167	1.28%
Medium Risk	610	0.512	±0.193	54.92%
High Risk	1015	0.785	±0.079	88.37%

The interpretability and clinical trustworthiness of the AI system are significantly improved with this risk stratification methodology. The Bayesian model grants more knowledgeable triage and resource distribution by giving not just a predicted probability but also a confidence estimates for each decision. In clinical settings were reducing false negatives and marking uncertain cases for follow-up are essential for better patient outcomes, this kind of framework is particularly important.

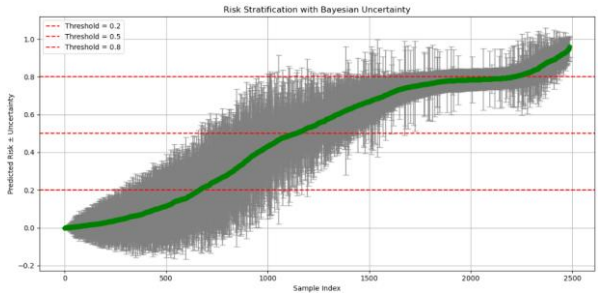

Figure 4: Risk stratification plot using Bayesian predictions: Cases categorized as low, medium, and high risk based on predicted probability.

4.6 Comparative ROC analysis of models

The ROC Curve presented in Figure 5 illustrates the performance of various learning models, namely Gradient Boosting (GBM), XGBoost, LightGBM, CatBoost, AdaBoost, and Stacked Model, for the detection of cardiovascular disease. The AUC values found under the corresponding curves represent the predictability of the model, where higher values imply better separation of cases into positive and negative. The resulting performance indicated that XGBoost AUC (0.9907), LightGBM AUC (0.9864), and CatBoost AUC (0.9818) were superior among the base learners for prediction, whereas Gradient Boosting AUC (0.9057) and AdaBoost AUC (0.7839) were inferior. The AUC of 0.9931 was the highest for the stacked model that outperformed every single base model. This result demonstrates the importance of stacked ensemble learning that incorporates the strengths of different classifiers for the increase of overall predictive accuracy and robustness. The predominance of the stacked model confirms that the combination of different learning algorithms results in better generalization and reliability, making it the most effective method for cardiovascular disease prediction. As shown in Figure 5, the ROC curve of the stacked ensemble model consistently dominates those of all individual base classifiers across the full range of classification thresholds. The stacked model achieves the highest AUC-ROC value of 0.9931, demonstrating superior discriminative ability between CVD-positive and CVD-negative cases. This performance highlights the effectiveness of combining heterogeneous learners through stacking, resulting in improved robustness and generalization compared to single-model approaches.

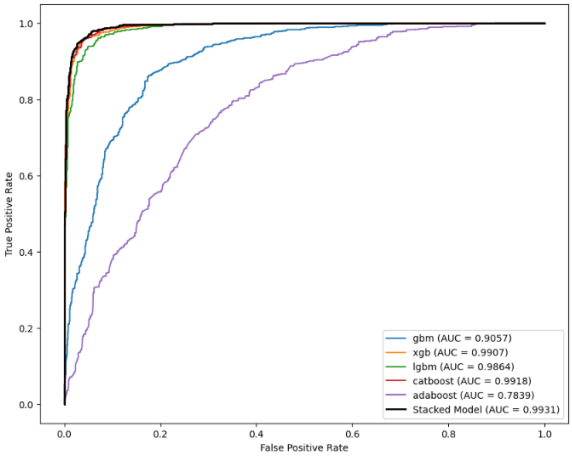

Figure 5: Receiver Operating Characteristic (ROC) curve comparing the classification performance of different models.

The reported results advance the field of cardiovascular risk prediction by demonstrating that the integration of temporal embeddings, stacked ensemble learning, and Bayesian uncertainty estimation leads to both superior predictive performance and improved clinical reliability. Beyond achieving higher accuracy and AUC-ROC values compared to existing models, the proposed framework introduces uncertainty-aware risk assessment, enabling more informed and safer clinical decision-making. These results highlight a shift from purely performance-driven models toward interpretable and confidence-aware AI systems suitable for real-world healthcare deployment.

5 Discussion

The study presents a new predictive framework for CVD risk prediction by integrating temporal feature extraction with LSTM autoencoders, feature aggregation, ensemble stacking learning, and Bayesian estimation of uncertainty. The framework overcomes three important challenges of current predictive models: lack of learning temporal dependencies, trouble with class imbalance, and lack of measurably expressible confidence with predictions. By combining stationary principal components with time-dynamic LSTM-based embeddings and utilising an ensemble meta-learning strategy with robustness, the new model outstrips previous methods both in accuracy and interpretability. The addition of Bayesian inference provides an additional level of clinical utility by providing probabilistic risk measures and ranges of uncertainty that are necessary for risk-conscious medical decision-making.

The proposed framework shows better performance and superior research methods when compared to current cardiovascular disease prediction models which represent their highest performance level. Previous research studies have focused on using unchanging data representations together with their associated algorithms which resulted in accuracy rates that ranged from 85 percent to 92 percent. The proposed model reached an accuracy score of 96.06 percent while achieving an AUC-ROC score of 99.31 percent because it used LSTM-based temporal embeddings and stacked ensemble learning and Bayesian uncertainty estimation. The proposed framework brings a novel approach which increases prediction accuracy and delivers risk assessment that includes uncertainty information for better clinical outcomes and decision-making assistance.

Although the Framingham Heart Study dataset is a widely used benchmark for cardiovascular risk modeling, it represents a specific population cohort and does not fully capture the ethnic, geographic, and socio-economic diversity observed in global clinical settings. Consequently, while the strong performance reported in this study demonstrates the methodological effectiveness of the proposed framework, it should not be interpreted as universal generalizability across all populations. External validation using large-scale, multi-ethnic, and multi-institutional datasets is essential to assess robustness and mitigate potential population-specific bias. Future work will therefore focus on evaluating the framework on diverse longitudinal cohorts to further establish its clinical applicability and fairness.

In contrast to standard classification pipelines that utilize only static features, the authors of this study took advantage of simulated temporal data which had been processed by an LSTM autoencoder. The technique used reveals the very complicated longitudinal patterns the heart's health indicators show, such as changes in systolic pressure, cholesterol, and blood glucose. The merger with principal components from the static clinical features produced a 41-dimensional latent space which was ripe for further learning. The dual representation kept both the structural and the time-changing dimensions of the patient profiles, which facilitated the extraction of more profound risk trajectory understanding. As a result of the enhancement in the feature space, the stacked ensemble model was able to learn more discriminative boundaries, which in turn led to better generalization across borderline and high-risk cases.

The architecture proposed represents a big leap compared to prior studies which mainly applied classical machine learning methods on handcrafted or univariate features. For example, Sudipta et al. (2022) reported an accuracy of 87.70% with a Multilayer Perceptron (MLP) over a range of datasets, such as Cleveland, Hungarian, Switzerland, Long Beach, and StatLog, while using infinite feature selection techniques [48]. Similarly, Reddy et al. (2021) used Sequential Minimal Optimization (SMO) on the Cleveland Heart Dataset, reaching 85.15% accuracy with the full attribute set and 86.47% with an optimized attribute selection [49]. Other studies employed different machine learning techniques but still recorded lower predictive performance. Mohan et al. (2019) introduced a Hybrid Random Forest with Linear Model (HRFLM) on the UCI Cleveland dataset, achieving an accuracy of 88.7% using 13 clinical features [50]. Ambrews et al. (2022) improved performance slightly with a Voting Ensemble Model applied to the UCI dataset, obtaining 91.96% accuracy [51]. The study by Mohapatra et al. (2023) employed a stacked ensemble model with ten base learners, including Random Forest (RF), MLP, KNN, Extra Trees (ET), XGBoost, Support Vector Classifier (SVC), Stochastic Gradient Descent (SGD), AdaBoost (ADB), CART, and Gradient Boosting Machine (GBM), and reached 92% accuracy [18].

One of the most attractive features of the framework is its capacity to provide clinically interpretable risk stratification via Bayesian modeling. The implementation of Monte Carlo Dropout allowed the Bayesian neural network to create the prediction distribution per test instance where mean probabilities and uncertainty intervals were extracted. As shown in the confidence interval plot (Figure 3), the model consistently exhibited great certainty for unambiguous very low and very high risk situations, whereas it revealed more uncertainty at the 0.5 decision threshold. This is in line with the clinical assumption that there has been a mixture or at least a vague overlapping of features in the case of borderline situations. More importantly, this characteristic allows doctors to choose the most uncertain cases for follow-up diagnostics, thus, making it a safety measure in those deployment scenarios where losing a negative case would be very costly.

The risk stratification plot (Figure 4) was a further confirmation of the model's capability to divide the population into clinically relevant categories. Patients were categorized automatically into risk tiers of Low (<0.33), Medium (0.33–0.66), and High (>0.66) based on Bayesian predicted probabilities. The actually measured CHD incidence rate in the group of patients classified as High Risk was 88.37%, thus confirming the total agreement between the predicted and observed outcomes. On the contrary, only 1.28% of the patients from the Low Risk group had the disease. Not only did these results confirm the predictive model's accuracy, but they also pointed to the model's potential use in the actual triaging of patients in clinical settings. The stratification also showed that the Medium Risk group had the highest uncertainty (±0.193), thus directing clinicians to monitor or reevaluate these patients with more diligence. The categorization based on risk awareness can help in making better decisions regarding the allocation of resources, patient counseling, and planning of proactive interventions.

The research, besides the other positives, dealt with the issue of class imbalance through a very constructive approach, which was another main benefit of it. The positive class (at-risk) makes up only 15.19% of the total samples, which makes the original dataset highly uneven. If this issue is not addressed, it will lead to the creation of biased models that will support the majority class at the cost of clinical safety. Employing Synthetic Minority Over-Sampling Technique (SMOTE) in the fused feature space not only helped the study in achieving class distribution that is balanced but also kept the dynamic and static characteristics of the features intact. This method drove the stacked ensemble and Bayesian models to outline universal decision boundaries for the entire dataset thus leading to a remarkable recalling of the minority class from 80.21% (in AdaBoost) to 97.67% in the final ensemble model. So a high recall is especially vital in medical diagnostics because false negatives may cause the delay of critical interventions or even their non-performance at all.

The individual base models' performance comparison comes to be a good example of ensemble learning value. AUCs above 98% marked the good performance of XGBoost, CatBoost, and LightGBM as individual models. However, each of them had a small trade-off in precision and recall. For instance, CatBoost had the highest F1-score among base learners (95.03%), but LightGBM had the best recall (97.75%) with a slight decline in precision. The ensemble model, by the use of a Multilayer Perceptron meta-learner for integrating predictions from these learners, combined their individual strengths while decreasing weaknesses and thus, a balanced performance profile was obtained across all key metrics. This synergy, which is a characteristic feature of well-designed ensemble architectures, was very important in taking the final model’s accuracy and robustness beyond that of any single learner.

The model's interpretability and clinical applicability were additionally improved with the help of an uncertainty-aware decision-support system. The lack of confidence estimates in conventional black-box models often hinders clinical acceptance. Conversely, the Bayesian method used in this research allows doctors to receive not only a risk probability but also an estimate of the model's confidence in that specific prediction. For example, as demonstrated in Figure 4 and Table 3, predictions with high confidence were associated with both extremely high and extremely low risk probabilities, while predictions with middle-range confidence were marked with larger standard deviations, indicating the need for further diagnostic input. This encourages a model-in-the-loop strategy where AI aids but does not take over clinical judgment thus the process being safer and more transparent through decision-making.

Nevertheless, there are limitations in the research, no matter the advancements. Temporal data was first synthetically simulated through replication which meant that it was not derived from longitudinal patient records and this was probably the main reason the model could not completely capture the temporal variability present in the real-world EHR systems. Therefore, integrating genuine longitudinal datasets for the detection of evolving risk profiles should be the future of the model's validation. The model used in this study was evaluated on the Framingham dataset, but still, external validation on multi-institutional and ethnically diverse populations is a must to confirm generalizability. The cardiovascular risk factors and disease prevalence differ significantly from one demographic group to another, and consequently, training on a more heterogeneous dataset would not only prevent bias but also contribute to the fairness of the deployment.

The use of synthetically constructed temporal sequences may also influence the reliability of the model when applied to real-world longitudinal data. While the replicated sequences enable the LSTM autoencoder to learn latent feature interactions within a sequential framework, they do not capture true temporal variability arising from lifestyle changes, disease progression, or treatment effects over time. When exposed to authentic longitudinal data, the model may exhibit different sensitivity to evolving risk patterns, potentially improving both predictive accuracy and uncertainty calibration. Consequently, the current results should be interpreted as demonstrating methodological feasibility rather than definitive longitudinal performance, and future validation on real-world time-series clinical data is essential to fully assess reliability in practical deployment settings.

Another aspect that researchers can explore in the future is the optimization of the model and its deployability. Though the present structure is precise, it still goes through several learning stages and has a moderately complicated fusion pipeline, which might restrict its use in terms of real-time or low-resource environments. Simplifying the structure, perhaps by combining dimensionality reduction with the LSTM network or looking into transformer-based temporal encoders, might lead to a reduction in complexity and a decrease in the demand for computational resources. Likewise, applying the Bayesian MLP in a lighter version on edge devices such as wearable monitors or mobile health applications would bring its usage closer to point-of-care especially in disadvantaged or rural areas.

Last but not least, the explainability of the model can be boosted even more by the application of global and local interpretability tools, which are not limited to the Bayesian confidence intervals. While the probabilistic nature of the current model offers some degree of transparency, doctors usually prefer to have graphical or textual interpretations of feature contributions concerning particular patients. Methods like SHAP (Shapley Additive Explanations) or counterfactual reasoning could go hand in hand with the present uncertainty plots and provide a more sophisticated comprehension of the factors that cause each prediction.

This situation would be ideal for patient education and shared decision-making, as it would help to cultivate trust in AI-assisted diagnosis.

The proposed framework advances the current state of cardiovascular risk prediction in three important ways. The research establishes temporal representation learning through LSTM autoencoders which enables the system to detect hidden inter-feature relationships without needing complete longitudinal data because most research treats patient information as unchanging static conditions. The stacked ensemble architecture with its trained meta-model system delivers improved performance through its ability to combine different classifier systems compared to the single-model and voting systems which studies in the literature typically use. The framework uses Bayesian uncertainty estimation to enhance its predictive capabilities because it enables the system to perform confidence-aware risk assessment which existing cardiovascular prediction models do not offer. The research presents a new method of cardiovascular risk assessment which enables medical professionals to make decisions based on clear results while understanding the uncertainty of their work which improves both clinical reliability and real-world implementation.

The study presents positive findings, but multiple limitations need to be recognized. First, Framingham Heart Study dataset exists as a specific population dataset which cannot represent all the ethnic, geographic, and socio-economic groups that exist in real-world clinical settings thus creating population bias which restricts general study applications. Second, researchers used synthetic time intervals created through feature duplication because actual patient records did not exist yet this method effectively supported representation learning while it failed to show actual disease development throughout time. The training data used SMOTE to address class imbalance problems present in cardiovascular datasets, but actual model performance will be affected by existing class imbalances when the model operates in environments with extreme class disparities. The hybrid framework requires multiple preprocessing and learning stages which results in increased computational expenses that create difficulties during deployment in real-time environments or systems with limited resources. The existing limitations require future research to test results on extensive longitudinal datasets from various institutions while researchers need to study fairness and robustness and deployment efficiency across different medical environments.

In a nutshell, this study delivers a model which is interpretable, technically valid and considering clinical relevance for predicting the risk of getting cardiovascular disease. By incorporating the temporal embeddings that LSTM learned along with ensemble learning and Bayesian confidence modeling, it reaches a state-of-the-art level concerning the solutions to major problems faced in medical AI practice. Through accuracy, stratified risk understanding, and predictive uncertainty estimation all working together, the end product of this model is a powerful tool for early-stage treatment and personalized care of cardiovascular disease. After validation, fine-tuning, and improving toward interpretability, this model will not be directly impactful but rather contribute indirectly through practical utility in the form of dependable, AI-based health technologies in clinical practice.

6 Conclusion

The study presents an advanced model for the assessment of cardiovascular disease (CVD) risk that combines temporal representation learning with LSTM autoencoders, static and dynamic health markers feature fusion, and ensemble classification with different gradient boosting classifiers, along with Bayesian inference-based uncertainty estimation. The proposed method is able to provide an extraordinary 96.06% accuracy and 99.31 AUC-ROC predictive power while it is also able to deal with the typical drawbacks associated with the CVD prediction models such as class imbalance, temporal unawareness, and absence of model confidences. The application of LSTM-based embeddings helped in recognizing changing risk factors, while the addition of PCA components through feature fusion expanded the representational space with increased discrimination ability. The structure-based stacking ensemble further contributed to the reliance by taking the cross-strengths between the classifiers of XGBoost, CatBoost, LightGBM, Gradient Boosting, and AdaBoost. Simultaneously, the Bayesian MLP model yielded clinically interpretable measures of uncertainty, hence it was very important to consider the clinical aspect of the threshold predictions very delicately. The analysis of risk classification validated the model's capability to classify the patients correctly into three groups of risk-low, medium, and high--which in turn facilitated more personalized and less ambiguous decision-making in healthcare. The future evaluation of the proposed framework will use large-scale datasets which contain data from multiple institutions and various ethnic groups to test its performance and fairness and its ability to function across different population groups. The study will investigate different temporal modeling architectures which include transformer-based encoders and lightweight recurrent networks to develop better cardiovascular risk pattern modeling methods that consume less computational resources. The study will explore two main research areas which include real-time integration with wearable and mobile health systems and the use of explainability techniques for uncertainty estimation to build clinical trust and enable practical implementation in preventive cardiology and population health screening and ongoing patient monitoring.

References

[1] A. S. Mohd Faizal, T. M. Thevarajah, S. M. Khor, and S.-W. Chang, “A review of risk prediction

models in cardiovascular disease: conventional approach vs. artificial intelligent approach," *Computer Methods and Programs in Biomedicine*, vol. 207, p. 106190, Aug. 2021, doi: 10.1016/j.cmpb.2021.106190.

[2] A. Rajdhan, A. Agarwal, M. Sai, D. Ravi, and D. P. Ghuli, "Heart Disease Prediction using Machine Learning," *International Journal of Engineering Research*, vol. 9, no. 04.

[3] M. Chiarito, L. Luceri, A. Oliva, G. Stefanini, and G. Condorelli, "Artificial Intelligence and Cardiovascular Risk Prediction: All That Glitters is not Gold.," *Eur Cardiol*, vol. 17, p. e29, Feb. 2022, doi: 10.15420/ecr.2022.11.

[4] K. Drożdż *et al.*, "Risk factors for cardiovascular disease in patients with metabolic-associated fatty liver disease: a machine learning approach," *Cardiovascular Diabetology*, vol. 21, no. 1, p. 240, Nov. 2022, doi: 10.1186/s12933-022-01672-9.

[5] C. Estes *et al.*, "Modeling NAFLD disease burden in China, France, Germany, Italy, Japan, Spain, United Kingdom, and United States for the period 2016–2030," *Journal of Hepatology*, vol. 69, no. 4, pp. 896–904, Oct. 2018, doi: 10.1016/j.jhep.2018.05.036.

[6] V. Shorewala, "Early detection of coronary heart disease using ensemble techniques," *Informatics in Medicine Unlocked*, vol. 26, p. 100655, Jan. 2021, doi: 10.1016/j.imu.2021.100655.

[7] J. Li, A. Loerbroks, H. Bosma, and P. Angerer, "Work stress and cardiovascular disease: a life course perspective," *Journal of Occupational Health*, vol. 58, no. 2, pp. 216–219, Mar. 2016, doi: 10.1539/joh.15-0326-OP.

[8] E. J. Benjamin *et al.*, "Heart Disease and Stroke Statistics—2019 Update: A Report From the American Heart Association," *Circulation*, vol. 139, no. 10, pp. e56–e528, Mar. 2019, doi: 10.1161/CIR.0000000000000659.

[9] H. S. N. Murthy and M. Meenakshi, "Dimensionality reduction using neuro-genetic approach for early prediction of coronary heart disease," in *International Conference on Circuits, Communication, Control and Computing*, Nov. 2014, pp. 329–332. doi: 10.1109/CIMCA.2014.7057817.

[10] Purushottam, K. Saxena, and R. Sharma, "Efficient Heart Disease Prediction System," *Procedia Computer Science*, vol. 85, pp. 962–969, Jan. 2016, doi: 10.1016/j.procs.2016.05.288.

[11] M. Jabbari *et al.*, "Development of a CVD mortality risk score using nutritional predictors: A risk prediction model in the Golestan Cohort Study," *Nutrition, Metabolism and Cardiovascular Diseases*, vol. 35, no. 1, p. 103770, Jan. 2025, doi: 10.1016/j.numecd.2024.10.008.

[12] X. Wan, X. Mei, Y. Chen, J. Luo, and L. Hao, "Automated arrhythmia classification based on a pyramid dense connectivity layer and BiLSTM," *Technology and Health Care*, vol. 33, no. 2, pp. 797–813, Mar. 2025, doi: 10.1177/09287329241290941.

[13] D. A. Anggoro, "Comparison of Accuracy Level of Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) Algorithms in Predicting Heart Disease," *IJETER*, vol. 8, no. 5, pp. 1689–1694, May 2020, doi: 10.30534/ijeter/2020/32852020.

[14] E. Canayaz, Z. A. Altikardes, A. Unsal, H. Korkmaz, and M. Gok, "Development and validation of machine learning algorithms for early detection of ankylosing spondylitis using magnetic resonance images," *Technology and Health Care*, p. 09287329241297887, Dec. 2024, doi: 10.1177/09287329241297887.

[15] J. Joseph and K. Kartheeban, "Visualizing the Full Spectrum Optimization of K-Nearest Neighbors From Data Preprocessing to Hyperparameter Tuning and K-Fold Validation for Cardiovascular Disease Prediction," *IJCAI*, vol. 49, no. 2, May 2025, doi: 10.31449/inf.v49i2.7774.

[16] B. Pavlyshenko, "Using Stacking Approaches for Machine Learning Models," in *2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)*, Aug. 2018, pp. 255–258. doi: 10.1109/DSMP.2018.8478522.

[17] A. Ghasemieh, A. Lloyed, P. Bahrami, P. Vajar, and R. Kashef, "A novel machine learning model with Stacking Ensemble Learner for predicting emergency readmission of heart-disease patients," *Decision Analytics Journal*, vol. 7, p. 100242, June 2023, doi: 10.1016/j.dajour.2023.100242.

[18] S. Mohapatra *et al.*, "A stacking classifiers model for detecting heart irregularities and predicting Cardiovascular Disease," *Healthcare Analytics*, vol. 3, p. 100133, Nov. 2023, doi: 10.1016/j.health.2022.100133.

[19] H. Yang and J. M. Garibaldi, "A hybrid model for automatic identification of risk factors for heart disease," *Journal of Biomedical Informatics*, vol. 58, pp. S171–S182, Dec. 2015, doi: 10.1016/j.jbi.2015.09.006.

[20] G. I. Choudhary and P. Fränti, "Predicting onset of disease progression using temporal disease occurrence networks," *International Journal of Medical Informatics*, vol. 175, p. 105068, July 2023, doi: 10.1016/j.ijmedinf.2023.105068.

[21] F. M. Alkoot, Hussain. M. Alkhedher, and Z. F. Alkoot, "Experimental analysis of machine learning methods to detect Covid-19 from x-rays," *Journal of*

Engineering Research, vol. 11, no. 2, p. 100063, June 2023, doi: 10.1016/j.jer.2023.100063.

[22] M. A. Almulla, "A multimodal emotion recognition system using deep convolution neural networks," *Journal of Engineering Research*, vol. 13, no. 2, pp. 721–729, June 2025, doi: 10.1016/j.jer.2024.03.021.

[23] W. J. Von Eschenbach, "Transparency and the Black Box Problem: Why We Do Not Trust AI," *Philos. Technol.*, vol. 34, no. 4, pp. 1607–1622, Dec. 2021, doi: 10.1007/s13347-021-00477-0.

[24] X. Zhou, B. Chen, Y. Gui, and L. Cheng, "Conformal Prediction: A Data Perspective," *ACM Comput. Surv.*, p. 3736575, May 2025, doi: 10.1145/3736575.

[25] T. J. Loftus *et al.*, "Uncertainty-aware deep learning in healthcare: A scoping review," *PLOS Digit Health*, vol. 1, no. 8, p. e0000085, Aug. 2022, doi: 10.1371/journal.pdig.0000085.

[26] J. Joseph and K. Kartheeban, "Exploring Missing Value Handling Techniques for Optimized KNN Heart Disease Prediction," in *2024 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES)*, Sept. 2024, pp. 1–8. doi: 10.1109/SPICES62143.2024.10779729.

[27] S. Gour, P. Panwar, D. Dwivedi, and C. Mali, "A Machine Learning Approach for Heart Attack Prediction," in *Intelligent Sustainable Systems*, A. K. Nagar, D. S. Jat, G. Marín-Raventós, and D. K. Mishra, Eds., Singapore: Springer Nature Singapore, 2022, pp. 741–747.

[28] C. Gupta, A. Saha, N. V. Subba Reddy, and U. Dinesh Acharya, "Cardiac Disease Prediction using Supervised Machine Learning Techniques.," *J. Phys.: Conf. Ser.*, vol. 2161, no. 1, p. 012013, Jan. 2022, doi: 10.1088/1742-6596/2161/1/012013.

[29] Y. Chen *et al.*, "Automated Alzheimer's disease classification using deep learning models with Soft-NMS and improved ResNet50 integration," *Journal of Radiation Research and Applied Sciences*, vol. 17, no. 1, p. 100782, Mar. 2024, doi: 10.1016/j.jrras.2023.100782.

[30] H. El-Sofany, B. Bouallegue, and Y. M. A. El-Latif, "A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method," *Scientific Reports*, vol. 14, no. 1, p. 23277, Oct. 2024, doi: 10.1038/s41598-024-74656-2.

[31] A. J. Almalki, "OVGGNet: Optimized deep learning for lesion segmentation of medical images using color features," *Journal of Radiation Research and Applied Sciences*, vol. 18, no. 3, p. 101592, Sept. 2025, doi: 10.1016/j.jrras.2025.101592.

[32] C. Krittanawong *et al.*, "Machine learning prediction in cardiovascular diseases: a meta-analysis," *Scientific Reports*, vol. 10, 2020, doi: 10.1038/s41598-020-72685-1.

[33] M. González-Del-Hoyo and X. Rossello, "Challenges and promises of machine learning-based risk prediction modelling in cardiovascular disease.," *Eur Heart J Acute Cardiovasc Care*, vol. 10, no. 8, pp. 866–868, Oct. 2021, doi: 10.1093/ehjacc/zuab074.

[34] V. Nagavallika, "Prediction of Heart Disease Using Machine Learning Techniques," vol. 4, no. 56, 2022.

[35] A. C. Dimopoulos *et al.*, "Machine learning methodologies versus cardiovascular risk scores, in predicting disease risk," *BMC Med Res Methodol*, vol. 18, no. 1, p. 179, Dec. 2018, doi: 10.1186/s12874-018-0644-1.

[36] Prof. Madhavi Tota, Manthan Moon, Pranit Nagrale, Akshay Pandav, and Gunjan Das, "Heart Diseases Prediction System using ML," *IJARSCT*, pp. 337–345, Dec. 2022, doi: 10.48175/IJARSCT-7798.

[37] B. Jin, C. Che, Z. Liu, S. Zhang, X. Yin, and X. Wei, "Predicting the Risk of Heart Failure With EHR Sequential Data Modeling," *IEEE Access*, vol. 6, pp. 9256–9261, 2018, doi: 10.1109/ACCESS.2017.2789324.

[38] A. S. S. Kotia, M. Rastogi, and R. A. Bhongade, "Use of machine learning techniques for effective prediction of heart disease," *CM*, no. 26, pp. 315–321, Mar. 2023, doi: 10.18137/cardiometry.2023.26.315321.

[39] C. M. Bhatt, P. Patel, T. Ghetia, and P. Mazzeo, "Effective Heart Disease Prediction Using Machine Learning Techniques," *Algorithms*, vol. 16, p. 88, 2023, doi: 10.3390/a16020088.

[40] D. Shah, S. Patel, and S. K. Bharti, "Heart Disease Prediction using Machine Learning Techniques," *SN Computer Science*, vol. 1, no. 6, p. 345, Oct. 2020, doi: 10.1007/s42979-020-00365-y.

[41] E. D. Adler *et al.*, "Improving risk prediction in heart failure using machine learning," *European J of Heart Fail*, vol. 22, no. 1, pp. 139–147, Jan. 2020, doi: 10.1002/ejhf.1628.

[42] I. M. Pires, G. Marques, N. M. Garcia, and V. Ponciano, "Machine learning for the evaluation of the presence of heart disease," *Procedia Computer Science*, vol. 177, pp. 432–437, 2020, doi: 10.1016/j.procs.2020.10.058.

[43] Yuda Syahidin, Aditya Pratama Ismail, and Fawwaz Nafis Siraj, "Application of Artificial Neural Network Algorithms to Heart Disease Prediction Models with Python Programming," *E-Komtek*, vol.

6, no. 2, pp. 292–302, Dec. 2022, doi: 10.37339/e-komtek.v6i2.932.

[44] Yichun Wang, "Heart disease prediction with discriminative deep neural network," presented at the Proc.SPIE, May 2023, p. 126401P. doi: 10.1117/12.2673756.

[45] B. F. Azevedo, A. M. A. C. Rocha, and A. I. Pereira, "Hybrid approaches to optimization and machine learning methods: a systematic literature review," *Mach Learn*, Jan. 2024, doi: 10.1007/s10994-023-06467-x.

[46] S. .S, S. Lavanya, M. R. Chandhini, R. Bharathi, and K. Madhulekha, "Hybrid Machine Learning Techniques for Heart Disease Prediction," *International Journal of Advanced Engineering Research and Science*, vol. 7, pp. 44–48, Jan. 2020, doi: 10.22161/ijaers.73.7.

[47] N. A. Rajendran and D. R. Vincent, "Heart disease prediction system using ensemble of machine learning algorithms," *Recent Patents on Engineering*, vol. 15, no. 2, pp. 130–139, 2021.

[48] M. Sudipta, E. Abdel-Raheem, and L. Rueda, *Heart Disease Prediction Using Adaptive Infinite Feature Selection and Deep Neural Networks*. 2022, p. 240. doi: 10.1109/ICAIIC54071.2022.9722652.

[49] K. V. Reddy, I. Elamvazuthi, A. A. Aziz, S. Paramasivam, H. N. Chua, and S. Pranavanand, "Heart Disease Risk Prediction Using Machine Learning Classifiers with Attribute Evaluators," *Applied Sciences*, vol. 11, no. 18, 2021, doi: 10.3390/app11188352.

[50] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[51] A. B. Ambrews, E. Gubin Moung, A. Farzamnia, F. Yahya, S. Omatu, and L. Angeline, "Ensemble Based Machine Learning Model for Heart Disease Prediction," in *2022 International Conference on Communications, Information, Electronic and Energy Systems (CIEES)*, Nov. 2022, pp. 1–6. doi: 10.1109/CIEES55704.2022.9990665.