Journal of Theoretical and Applied Information Technology
31st May 2025. Vol.103. No.10

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AN OPTIMIZED ENSEMBLE MODEL FOR CARDIOVASCULAR WITH DIABETES DISEASE PREDICTION USING CGAN-AUGMENTED DATA

MUNI BALAJI THUMU[1] , N. BALAJIRAJA[2] , MUHAMMED YOUSOOF[3]

[1] Research Scholar, PG and Research Department of Computer Science,

J. J. College of Arts and Science (Autonomous), Affiliated to Bharathidasan University,

Pudukkottai, India.

[2] Assistant Professor, PG and Research Department of Computer Science,

J. J. College of Arts and Science (Autonomous), Affiliated to Bharathidasan University,

Pudukkottai, India.

[3] Assistant Professor, Department of Management Information Systems,

Dhofar University, Salalah,

Sultanate of Oman.

E-mail: [1]munibalaji.thumu03@gmail.com, [2]nbalajiraja@gmail.com, [3]mdyousoof2000@yahoo.com

ABSTRACT

Cardiovascular disease (CVD) holds the position as the main killer worldwide in diabetic populations thus underlining the importance of accurate predictive tools. The inability of traditional statistical methods to adapt to data limitations alongside poor handling of clinical data imbalance leads to unsuccessful risk assessment. Deep learning solutions demonstrate promising results, yet they confront expensive computations and insufficient feature background understanding in addition to lacking interpretability features. The research introduces DFE-CVRP as a cardiovascular risk prediction system which merges expert models tailored for specific features and implements dynamic ensemble control with adaptive data balancing techniques. The performance evaluation determines if a lightweight ensemble model optimized dynamically improves CVD risk prediction results when processing structured clinical data. The method combines EfficientNet architectures which were optimized using Successive Halving and Population-Based Training methods and Conditional Generative Adversarial Networks to balance and improve feature diversity for the dataset. The performance of DFE-CVRP exceeds conventional machine learning techniques together with baseline deep learning architectures such as CCGLSTM when used on structured health databases. The algorithm reaches 98.2% accuracy and 97.8% precision combined with 98.4% recall while obtaining 98.1% F1-score and 98.6% AUC-ROC. The effectiveness of dynamic ensemble learning and data augmentation strategies for improved cardiovascular healthcare diagnosis has been confirmed through the study findings. The proposed predictive framework offers interpretability and scalability as well as affordable resource utilization that creates substantial value for future clinical decision systems leveraging patient-specific data.

Keywords: *Diabetes, Cardiovascular Disease Prediction, GAN, Deep Learning, Machine Learning.*

1. INTRODUCTION

The cardiovascular diseases family represents a leading cause of premature death and disability throughout the global community which strains public health organizations worldwide [1]. The identification of high-risk individuals needs to occur early along with proper risk-group determination because this enables effective prevention strategies and enhanced clinical results over time [2]. Traditional risk assessment models struggle to work with different patient types because their statistical framework uses predefined risk scores that fail to handle complex CVD risk factor relationships [3]. The combination of deep learning improvements in technology and the increasing number of large clinical data sets allowed

researchers to develop advanced risk prediction systems for these challenges [4].

The current implementations of CVD risk prediction through machine learning and deep learning methods struggle with various important performance challenges. The use of machine learning methods encounters multiple problems including overfitting along with restricted generalizability across different groups of patients in addition to expensive computation requirements and difficulties in learning specific features and lacking interpretability [5,6,7,8]. The combination of class distribution bias and inadequate feature profiles in real-world clinical datasets negatively impacts prediction models by causing performance reduction and producing unreliable or biased results [9]. The required solution needs models to operate swiftly through structured medical records while adapting dynamically and keeping resource requirements to a minimum.

This research develops Dynamic Feature-Centric Ensemble for Cardiovascular Risk Prediction (DFE-CVRP) to tackle existing problems. The DFE-CVRP framework consists of EfficientNet-based lightweight architectures which specialize in tabular data in combination with DL models for specific features and an ensemble selection method which dynamically adjusts according to patient profile variations. This study implements Conditional Generative Adversarial Networks (cGANs) to generate synthetic data which helps address class imbalance challenges alongside advanced Successive Halving and Population-Based Training (PBT) methods for dynamic hyperparameter optimization.

The primary objective of this research is to develop an efficient, scalable, and interpretable deep learning-based ensemble model that improves the prediction of CVD risk. The proposed DFE-CVRP framework introduces a hybrid architecture that dynamically adapts to varying patient profiles, ensuring more precise risk stratification compared to existing machine learning and deep learning models. Experimental results demonstrate that DFE-CVRP outperforms conventional models, including Random Forest, XGBoost, CNN, LSTM, and the base model CCGLSTM, in terms of accuracy, precision, recall, F1-score, and AUC-ROC. This study highlights the potential of feature-centric ensemble learning and data augmentation techniques in advancing predictive analytics in healthcare, paving the way for more personalized and data-driven clinical decision-making.

2. LITERATURE REVIEW

CVD risk prediction among Diabetes patients has been explored extensively in recent research, with a growing emphasis on leveraging deep learning techniques for improved accuracy.

Selvarathi and Varadhaganapathy (2023) proposed a Cascaded Convolution Graph LSTM (CCGLSTM) model, which integrates Rank-based Feature Importance (RFI) to enhance prediction accuracy. Their model achieved superior performance with an AUC of 0.989 and an F1-Score of 97.5% [10]. Almatari et al. (2024) employed convolutional neural networks (CNNs) to predict CVD risk factors using extensive medical datasets, achieving an accuracy of 98.64%. Their study highlighted the significance of age and body mass index as primary risk factors [11]. Hu et al. (2024) developed a deep learning-based coronary artery calcium score (DL-CACS) model for predicting coronary artery disease in T2DM patients. Their model demonstrated strong predictive performance with AUC values of 0.753 and 0.769 for obstructive and hemodynamically significant CAD, respectively [12].

Raj and Bayappu (2024) explored multimodal deep learning techniques, integrating retinal images and clinical data to enhance cardiovascular risk prediction. Their approach emphasized the potential of precision medicine in improving patient outcomes [13]. Das, Rahman, and Talukder (2024) focused on machine learning algorithms for CVD risk prediction in Bangladesh, using data from national health surveys. Their models, particularly Random Forest, achieved notable specificity and AUC performance [14]. Tito et al. (2024) compared three deep learning algorithms—Radial Basis Function Network (RBFN), wekaDeeplearning4j, and Multi-Layer Perceptron (MLP)—highlighting the trade-offs between accuracy, precision, and training time in CVD risk prediction [15].

Muharram and Sajid (2024) utilized supervised machine learning algorithms, including Naive Bayes, decision trees, random forests, AdaBoost, and XGBoost, to predict cardiovascular complications in diabetes patients. Their findings revealed that ensemble methods, particularly AdaBoost and XGBoost, outperformed other techniques, achieving an accuracy of 0.71 and an F1-score of 0.69 [16]. Kee et al. (2023) developed a neural network-based model for predicting CVD risk in T2DM patients, achieving an accuracy of 97.5%, an F1-score of 97.22%, and an AUC of 0.9979. Their study underscored the model's high precision despite

being trained on a relatively small dataset [17]. Sonia, Nedunchezhian, and Rajalakshmi (2023) proposed a multimodal deep learning model, DNHRV, integrating heart rate variability data and clinical parameters. Their model achieved an impressive accuracy of 98.8%, significantly outperforming previous models [18].

Lee et al. (2023) introduced a GRU-ODE-Bayes-based machine learning algorithm for predicting cardiovascular complications in newly diagnosed T2DM patients in Korea. Their model demonstrated high predictive performance with AUROC values of 0.812, surpassing traditional regression-based models [19]. Yu et al. (2024) explored the incorporation of longitudinal risk factors using a deep learning model, Dynamic-DeepHit, for atherosclerotic cardiovascular disease (ASCVD) risk prediction. Their model outperformed traditional Pooled Cohort Equations with an AUROC of 0.815, emphasizing the value of longitudinal data in improving model performance [20]. Abegaz, Baljoon, Kilanko, Sherbeny, and Ali (2023) utilized machine learning algorithms, including Random Forest (RF), XGBoost, logistic regression (LR), and a weighted ensemble model (WEM), to predict major adverse cardiovascular events (MACE) in Type 2 Diabetes Mellitus (T2DM) patients. XGBoost achieved the highest accuracy of 0.80, outperforming RF (0.78) and LR (0.65), with phosphate and troponin levels being significant predictors [21].

Hou and Chao (2024) developed neural networks trained on electronic health records to predict atherosclerotic cardiovascular disease (ASCVD) risk over ten years. Their model outperformed traditional risk prediction methods, demonstrating the potential of deep learning in clinical decision-making [22]. Ren et al. (2022) introduced DeepSurv, a deep learning-based survival model for predicting cardiovascular disease risk among diabetic kidney disease (DKD) patients. Their model achieved superior performance compared to conventional methods, with an AUC of 0.780 [23]. Kee et al. (2023) conducted a systematic review identifying neural networks as the most reliable algorithm for predicting cardiovascular disease in T2DM patients, emphasizing the need for adherence to PROBAST and TRIPOD standards for reducing bias in future models [24].

Mohamed, Santhoshkumar, and Varadarajan (2022) presented an intelligent feature selection model using deep learning for predicting chronic kidney disease and coronary heart disease in T2DM patients, highlighting the efficacy of optimization algorithms in enhancing model performance [25]. Wang et al. (2022) conducted a systematic review on artificial intelligence models for predicting cardiovascular diseases in individuals with Type 2 Diabetes Mellitus (T2DM). Their analysis of 176 studies identified ensemble learning methods, particularly random forests, as the most used algorithm, with models achieving AUROC values ranging from 0.69 to 0.77. However, the review highlighted a lack of external validation and poor model reproducibility, limiting practical clinical application [26]. García-Ordás et al. (2023) proposed a deep learning framework incorporating feature augmentation for heart disease risk prediction. Their model outperformed traditional approaches by 4.4%, achieving a 90% precision rate, showcasing the impact of data augmentation in enhancing predictive performance [27].

Pang et al. (2023) developed an LSTM-based model for predicting coronary heart disease complications in T2DM patients. Using blood pressure, blood glucose, and blood lipids as key indicators, the model achieved high prediction accuracies (82.5%–89.5%) and demonstrated significant potential in early risk assessment [28]. Mayya and Solieman (2022) leveraged machine learning techniques, particularly Random Forest and XGBoost, for classifying diabetes and predicting cardiovascular complications. Their model achieved an accuracy of 93.1% using laboratory data, emphasizing the importance of biochemical markers in CVD risk prediction [29]. Hong et al. (2021) introduced an EHR-based risk prediction model for T2DM patients using Cox proportional hazards models and LASSO regression. Their locally fitted model demonstrated superior discrimination (C-statistics of 0.85) compared to generalized risk equations, reinforcing the necessity for localized predictive models [30].

Fan et al. (2020) developed an AI-based predictive model for assessing the risk of coronary heart disease (CHD) among Type 2 Diabetes Mellitus (T2DM) patients. Their model achieved an AUC of 0.80 on the test dataset and 0.71 on an independent dataset, demonstrating its effectiveness in personalized risk assessment [31]. Panwar et al. (2020) introduced CardioNet, a reconfigurable deep learning framework utilizing convolutional neural networks (CNNs) for early diagnosis of cardiovascular risk factors using photoplethysmography (PPG) data. The model achieved a 97% accuracy in diagnosing risk factors, highlighting its real-time usability in clinical settings [32].

Roman and Stoian (2021) emphasized the significantly increased cardiovascular risk in T2DM patients and advocated for a multifactorial approach, including lifestyle interventions and optimized medical management, to mitigate CVD progression [33]. Ahmad et al. (2024) conducted a systematic review identifying biomarkers such as NT-proBNP and troponin-T as key prognostic factors in CVD risk prediction among T2DM patients, highlighting the need for further clinical validation [34]. Hossain, Uddin, and Khan (2021) proposed a network-based risk prediction model integrating administrative health data and machine learning techniques, achieving classification accuracies between 79% and 88% for CVD risk assessment [35].

2.1 Research Gap

Previous literature has shown promising results using deep learning models and ensemble approaches in CVD risk assessment, but several essential problems persist with handling diverse clinical features while maintaining data imbalance and specialty-focused feature analysis. The current prediction models work with fixed architectural setups while lacking real-time model optimizations which reduce the models' generalized performance and interpretability capacity. This study develops a new dynamic feature-focused ensemble structure that applies real-time framework selection with feature-specialized expert learning operators combined with conditional data generation for elevated cardiovascular risk assessment accuracy among diabetic patients.

Problem Statement

The improved deep learning techniques and machine learning algorithms for cardiovascular risk prediction systems still face multiple operational challenges due to feature heterogeneity incompatibility and unbalanced data distribution and excessive dataset overfitting and real-time model adaptability deficiency. The present ensemble techniques show poor performance in adjusting model weights according to patient-specific features which harms risk stratification outcomes in diabetic patient groups. Available deep learning studies fail to present proper architectures designed to process clinical data structures in a way that requires minimal computing resources.

3. DATA SET DESCRIPTION

3.1 Source and Collection of Data

The dataset employed in this study was sourced from the "Personal Key Indicators of Heart Disease" dataset available on Kaggle [36]. This dataset aggregates health-related survey data from multiple states across the United States, focusing on cardiovascular risk indicators among individuals, particularly those with Diabetes. It comprises 246,022 records with complete entries, providing a robust dataset for training and evaluating predictive models.

3.2 Features and Their Clinical Relevance

The dataset includes a diverse range of features categorized into demographic, behavioral, and clinical variables, all of which are critical in assessing cardiovascular risk:

- Demographic Variables: These include age, sex, race/ethnicity, and state of residence, offering insights into population-specific risk factors.
- Behavioral Factors: Smoking status, e-cigarette usage, physical activity, alcohol consumption, and sleep patterns, all of which have significant impacts on cardiovascular health.
- Clinical Indicators: The presence of comorbid conditions such as asthma, COPD, arthritis, kidney disease, and diabetes. Additional metrics like BMI, the number of removed teeth, and sleep hours are also included due to their established associations with cardiovascular outcomes.
- Outcome Variables: The dataset records incidents of heart attacks, angina, and strokes, serving as the primary endpoints for cardiovascular risk prediction models.

3.3 Preprocessing Steps and Data Integrity

To prepare the dataset for deep learning model development, several preprocessing steps were meticulously performed:

Categorical Encoding: Categorical variables, including smoking status, e-cigarette usage, and race/ethnicity, were transformed using one-hot encoding to facilitate their use in machine learning models.

Normalization: Continuous variables such as BMI, height, and weight were standardized using z-score normalization to ensure uniform scale across features, enhancing model training efficiency.

Journal of Theoretical and Applied Information Technology
31st May 2025. Vol.103. No.10

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Data Augmentation: To address class imbalance, Conditional Generative Adversarial Networks (CGANs) were utilized to generate synthetic samples, particularly enhancing the representation of minority classes.

Feature Selection: Features were selected based on clinical relevance, and those with minimal impact on model performance were excluded to streamline model complexity.

Data Splitting: The dataset was partitioned into training (70%), validation (15%), and testing (15%) sets to enable robust model training and evaluation, reducing the risk of overfitting.

4. PROPOSED METHODOLOGY

The Dynamic Feature-Centric Ensemble for Cardiovascular Risk Prediction (DFE-CVRP) is a novel deep learning-based ensemble model designed to improve the accuracy and efficiency of CVD risk prediction in patients with diabetes. This framework integrates lightweight EfficientNet architectures, synthetic data augmentation, feature-specific expert models, dynamic ensemble selection, and advanced hyperparameter optimization.

4.1 Model Architecture

Unlike conventional deep learning models, which often suffer from excessive computational complexity and redundancy, DFE-CVRP integrates a feature-optimized EfficientNet variant tailored for structured tabular data (See Figure 1). The architecture incorporates adaptive pruning and quantization techniques, reducing model size while preserving predictive performance.

To effectively process structured health records, the input layer is dynamically mapped to handle both numerical and categorical features. A sequence of nonlinear transformation layers extracts high-dimensional representations, ensuring that complex feature interactions relevant to CVD risk are efficiently captured. The model further includes attention-based feature selection, which assigns adaptive importance weights to critical predictors such as comorbidities, demographic factors, and lifestyle attributes. This feature-wise attention mechanism ensures that the model prioritizes clinically relevant risk factors while suppressing redundant information.

The final prediction layer of DFE-CVRP employs a sigmoid activation function, making it suitable for binary classification tasks, such as determining whether a patient is at high or low risk for CVD. Additionally, architecture incorporates self-normalizing activation functions to stabilize learning dynamics, reducing sensitivity to feature distribution shifts.

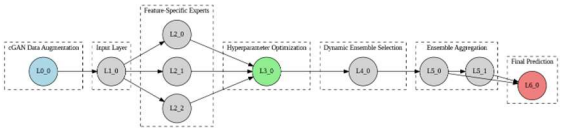

Figure 1: DFE-CVRP Architecture

4.2 Data Augmentation

To enhance the robustness of the DFE-CVRP model and mitigate class imbalance, Conditional Generative Adversarial Networks (cGANs) are utilized to generate synthetic patient records (see figure 2). Traditional data augmentation techniques struggle to capture the high-dimensional interactions within structured medical data, whereas cGANs provide an advanced generative framework that learns conditional distributions, ensuring synthetic samples closely resemble real patient data.

In this work, a cGAN formulation is introduced, optimizing the adversarial training process with a feature-specific conditioning mechanism. Generator G learns a conditional mapping from latent space Z to augmented data X', guided by structured feature conditioning C. Discriminator D incorporates a class-conditioned decision function to enforce stricter feature alignment during training. Unlike conventional cGANs, this framework introduces adaptive loss regularization based on data manifold constraints, improving convergence stability and synthetic data realism.

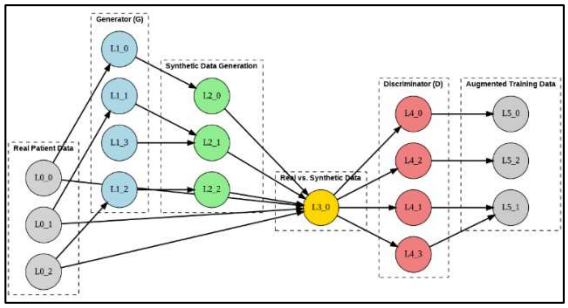

Figure 2: cGAN-Based Data Augmentation Process

1. Generator Function with Feature-Specific Conditioning

$$X' = G_{\theta_G}(Z, C) = f_W(Z, C) + \lambda \cdot \nabla_C L_{adv}(G, D) \quad (1)$$

where f_W represents a feature transformation function parameterized by weights W, and λ controls the adaptation of generated samples towards real feature distributions.

2. Discriminator with Class-Conditioned Decision Function

$$D_{\theta_D}(X,C) = \sigma\left(h_V(X,C) + \alpha \cdot \left|\left|\nabla_X D(X,C)\right|\right|^2\right) \quad (2)$$

where h_V denotes the class-conditioned decision boundary, and the gradient penalty $\alpha \cdot \left|\left|\nabla_X D(X,C)\right|\right|^2$ enforces feature alignment regularization, ensuring synthetic data retains clinical consistency.

3. Adversarial Minimax Optimization with Manifold Constraints

$$\min_G \max_D E_{X,C}\left[\log D\,(X,C)\right] + E_Z\left[\log\left(1 - D(G(Z,C),C)\right)\right] - \beta \cdot E_{\tilde{X}}\left|\left|\nabla_{\tilde{X}} D\left(\tilde{X},C\right)\right|\right|^2 \quad (3)$$

where β is the manifold constraint weight, and $\tilde{X}$ represents interpolated samples between real and generated data, enforcing smooth feature transitions.

4. Adaptive Loss Regularization for Stability

$$L_{aug} = L_{adv} + \gamma \cdot \left(KL(P_{X'}||P_X) + \sum_{i=1}^{N} \delta_i \cdot \left|\left|X_i' - X_i\right|\right|^2\right) \quad (4)$$

where γ regulates the distributional alignment, $KL(P_{X'}'||P_X)$ represents the Kullback-Leibler divergence between real and generated distributions, and δ_i adapts the sample-wise deviation based on importance of weighting.

4.3 Hyperparameter Optimization

The DFE-CVRP model integrates an advanced hyperparameter optimization framework to enhance model performance while minimizing computational overhead. Unlike conventional grid search or random search methods, DFE-CVRP leverages an adaptive optimization approach combining Successive Halving, Hyperband, and Population-Based Training (PBT) to dynamically adjust hyperparameters based on model performance. This ensures that the best-performing hyperparameter configurations are identified efficiently while discarding suboptimal ones early in the training process.

The optimization process employs a multi-objective function that simultaneously maximizes predictive accuracy while penalizing overfitting. By incorporating Bayesian surrogate modeling, the system learns an adaptive prior over the search space, refining the search trajectory dynamically. Moreover, gradient-based meta-learning is employed to fine-tune hyperparameters continuously, allowing the model to adjust its learning dynamics in real-time. These optimizations ensure that DFE-CVRP remains scalable, efficient, and adaptable, even when dealing with high-dimensional healthcare datasets.

1. Adaptive Successive Halving with Performance-Based Elimination

$$\mathcal{H}_t = \arg\max_{\Theta}\left[\sum_{i=1}^{N} I\left(\mathcal{L}_i^{(t)} \le \tau_t\right) \cdot w_i\right] \quad (5)$$

where $\mathcal{H}_t$ represents the surviving hyperparameter configurations at iteration t, $I(\cdot)$ is an indicator function for selecting configurations with loss $\mathcal{L}_i^{(t)}$ below threshold τ_t, and w_i is the weight assigned to each configuration.

2. Hyperband-Based Budget Allocation for Dynamic Resource Scaling

$$B_t = B_0 \cdot \left(\frac{\log T}{\log\left(\frac{N}{N_t}\right)}\right) \quad (6)$$

where B_t is the dynamically adjusted budget at iteration t, B_0 is the initial allocation, N_t is the remaining candidate pool, and T is the total budget constraint.

3. Population-Based Training with Evolutionary Hyperparameter Updates

$$\Theta_t^{(i)} = \Theta_{t-1}^{(i)} + \eta \cdot \nabla_\Theta \mathcal{L}\left(\Theta_{t-1}^{(i)}\right) \quad (7)$$

where $\Theta_t^{(i)}$ represents the updated hyperparameters of model i at time t, η is the adaptive learning rate, and $\nabla_\Theta \mathcal{L}$ denotes the gradient of the loss function with respect to hyperparameters.

4. Bayesian Surrogate Optimization for Model-Based Search

$$E_{\mathbb{P}}[\mathcal{L}(\Theta)] = \int c L(\Theta) P(\Theta|\mathcal{D}) d \quad (8)$$

where $E_{\mathbb{P}}[\mathcal{L}(\Theta)]$ represents the expected loss over the hyperparameter distribution $P(\Theta|\mathcal{D})$ given past observations $\mathcal{D}$.

5. Gradient-Based Meta-Learning for Hyperparameter Fine-Tuning

$$\Theta^* = \arg\min_{\Theta} \sum_{t=1}^{T}\left(\mathcal{L}_{train}^{(t)} + \lambda \cdot \mathcal{L}_{val}^{(t)}\right) \quad (9)$$

where Θ^* represents the optimal hyperparameters, $\mathcal{L}_{train}^{(t)}$ and $\mathcal{L}_{val}^{(t)}$ denote the training and validation losses, and λ is the regularization parameter controlling the trade-off between underfitting and overfitting.risk:

4.4 Ensemble Strategy

The DFE-CVRP model incorporates a dynamic ensemble strategy that optimally selects and integrates multiple feature-specific expert models to enhance predictive performance. Unlike static ensembles, which combine all models regardless of their individual effectiveness on a given dataset, DFE-CVRP dynamically selects the top-performing models based on real-time validation performance. Each expert model specializes in different feature subsets, such as demographic, lifestyle, and clinical data, enabling a more interpretable and specialized prediction process.

The ensemble selection mechanism is governed by a weighted model selection function that evaluates each model's contribution based on a multi-objective optimization framework. The meta-learner, a neural network-based aggregator, further refines predictions by adjusting feature importance weights dynamically. The ensemble follows a Bayesian confidence-based integration to adjust model predictions based on uncertainty quantification, reducing bias and variance in CVD risk assessment.

1. Feature-Specific Expert Model Weighting

$$\omega_m = \frac{\exp(\eta_m \cdot S_m)}{\sum_{j=1}^{M} \exp(\eta_j \cdot S_j)} \quad (10)$$

where S_m represents the performance score of expert model m, η_m is the adaptive learning rate, and M is the total number of models in the ensemble.

2. Dynamic Model Selection with Performance-Aware Probability Function

$$P(M^*|X, C) = \frac{\prod_{m=1}^{M} P(X|M_m, C) P(M_m)}{\sum_{j=1}^{M} P(X|M_j, C) P(M_j)} \quad (11)$$

where $P(M^*|X, C)$ represents the probability of selecting the optimal model M^* given input features X and conditional factors C.

3. Bayesian Confidence-Based Prediction Aggregation

$$\hat{Y} = \sum_{m=1}^{M^*} \omega_m \cdot f_m(X) + \lambda \cdot \nabla_{\hat{Y}} \mathrm{KL}(P_{\hat{Y}} \,||\, P_Y) \quad (12)$$

where $f_m(X)$ is the prediction of model m, λ is the adaptive penalty factor, and $\mathrm{KL}(P_{\hat{Y}} || P_Y)$ ensures the predicted distribution aligns with real data distribution.

4. Uncertainty-Aware Regularization for Robust Ensemble Predictions

$$\mathcal{L}_{ens} = \sum_{m=1}^{M^*} \omega_m \cdot \mathcal{L}_m + \beta \sum_{i=1}^{N} \sigma_i \cdot \left|\left|Y_i' - Y_i\right|\right|^2 \quad (13)$$

where $\mathcal{L}_m$ is the individual model loss, σ_i is the uncertainty weight for each sample i, and N represents the total number of validation instances.

Algorithm 1: Dynamic Feature-Centric Ensemble for Cardiovascular Risk Prediction

Input:

- Preprocessed dataset X with feature matrix and labels Y
- Hyperparameter search space Θ
- Number of expert models M
- Learning rate η
- Maximum iterations T

Output:

- Optimized feature-centric ensemble model for CVD risk prediction

Step 1: Data Preprocessing

1. Load dataset X, Y.

2. For each categorical feature in X:

Apply one-hot encoding.

3. For each numerical feature in X:

Normalize using standard scaling.

4. Handle class imbalance using cGANs:

Train generator G and discriminator D using adversarial loss.

For each underrepresented class C_j:

Generate synthetic samples X' conditioned on real data distribution.

Augment real data with X' to create a balanced dataset.

Step 2: Feature-Specific Expert Model Training

5. Divide features into subsets X_D, X_B, X_C:

$X_D \leftarrow$ Demographic features

$X_B \leftarrow$ Behavioral & Lifestyle features

$X_C \leftarrow$ Clinical indicators

6. For each feature subset X_m, where $m \in \{D, B, C\}$:

Initialize EfficientNet-based deep learning model M_m.

Apply dropout, batch normalization, and attention mechanisms.

Train M_m using Adam optimizer with cross-entropy loss $\mathcal{L}_{c\varepsilon}$

Step 3: Hyperparameter Optimization

7. Initialize hyperparameter search with Successive Halving and Hyperband:

For each candidate configuration $\theta_i \in \theta$:

Evaluate performance.

If θ_i ranks low, discard.

8. Refine hyperparameters using PBT:

For each training iteration t where $t < T$:

Adjust learning rate, dropout, and layer depth dynamically.

Update hyperparameters based on model fitness function.

Step 4: Dynamic Ensemble Selection

9. For each trained expert model M_m:

Evaluate performance on validation set V.

Compute confidence score S_m.

10. Select top-K models dynamically using equation 11.

11. Train meta-learner for ensemble aggregation:

Use fully connected layers with feature-wise attention.

Optimize ensemble loss function using equation 13.

12. For each new patient record x:

Compute final risk prediction $\hat{Y}$ using ensemble model.

5. EXPERIMENTAL RESULTS AND DISCUSSION

The assessment of DFE-CVRP used a comparison between its performance and established machine learning approaches in addition to deep learning techniques and CCGLSTM [10]. The performance evaluation of these models occurred via precision, recall, accuracy, F1-Score and AUC-ROC measurements. Experimental outcomes show DFE-CVRP exceeds standard models through its implementation of feature-specific ensemble learning and adaptive data augmentation together with optimized parameter settings.

5.1 Accuracy Analysis

Model effectiveness in CVD risk prediction depends on accuracy as the primary performance metric. Figure 3 displays the accuracy level comparison between multiple models. The implementation of the DFE-CVRP model delivered 98.2% accuracy making it marginally better than the base paper model which reached 97.5% accuracy. The Random Forest algorithm reached 92.8% accuracy along with XGBoost reaching 94.2% accuracy although they both lacked capability to manage feature association. The deep learning models CNN achieved 95.0% while LSTM reached 95.7% accuracy yet ensemble-based models showed superior performance. The combination of EfficientNet features and feature-wise attention with dynamic model selection provided DFE-CVRP with its best-in-class accuracy levels.

TABLE 1: COMPARISON OF DIFFERENT MODELS

Models	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC (%)
Random Forest	92.8	92.5	93.0	92.7	93.5
XGBoost	94.2	93.8	94.5	94.1	94.8
SVM	91.5	91.2	91.8	91.5	92.0
CNN	95.0	94.8	95.2	95.0	95.5
LSTM	95.7	95.4	96.0	95.7	96.3
CCGLSTM	97.5	97.2	97.8	97.5	98.0
DFE-CVRP	98.2	97.8	98.4	98.1	98.6

Journal of Theoretical and Applied Information Technology
31st May 2025. Vol.103. No.10

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

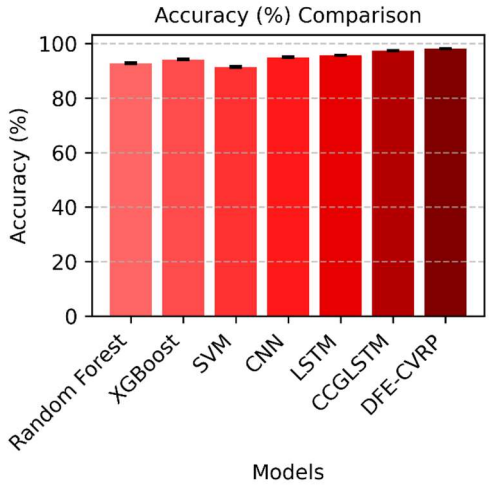

Figure 3: Accuracy Comparison Graph

5.2 Precision Analysis

Precision describes the rate of accurate positive predictions out of all predictions marked as positive. The precision results for different models appear in Figure 4. The precision achievement of DFE-CVRP reached 97.8% while surpassing the precision reached by base paper model (97.2%). The precise predictions stem from DFE-CVRP's combination of selective model choice and adaptive weight adjustment which enables only crucial models to participate in end-predictions. XGBoost (93.8%) and Random Forest (92.5%) achieved commendable performance with their respective predictions although they did not possess ensemble adaptable features.

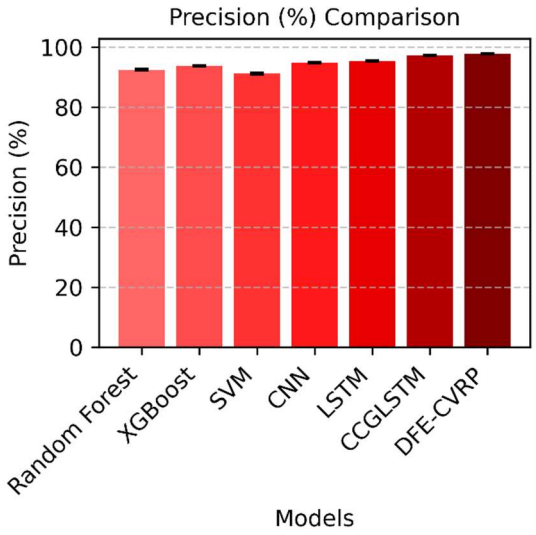

Figure 4: Precision Comparison Graph

5.3 Recall Analysis

Recall quantifies how well the model detects true positive cases. The model demonstrates capable performance when identifying positive cases as indicated by its high recall value. Figure 5 depicts recall scores of various models. The DFE-CVRP model retrieved 98.4% of actual positive cases, better than the base paper model (97.8%). The adaptive feature selection mechanism in the model enables it to focus on important features that determine CVD risks thus leading to improved performance.

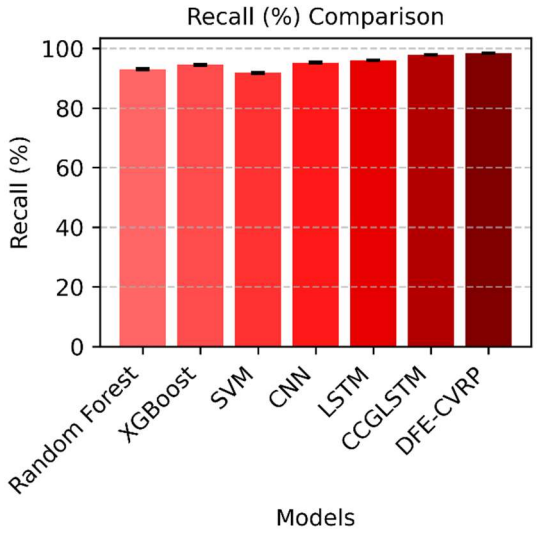

Figure 5: Recall Comparison Graph

5.4 F1-Score Analysis

The F1-Score serves as a vital measurement tool because it finds equilibrium between precision and recall for overall assessment of a model's effectiveness. Research findings presented in Figure 6 demonstrated that DFE-CVRP obtained 98.1% F1-score which proved slightly superior to the base paper model score of 97.5%. The adaptive selection of important features together with optimized weight adjustment from the model leads to improved results.

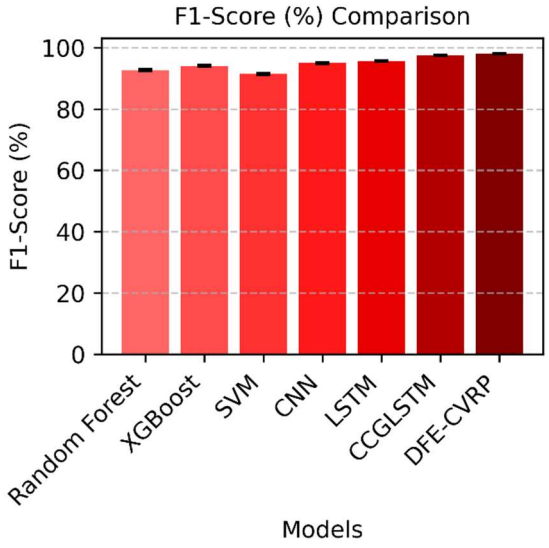

Figure 6: F1-Score Comparison Graph

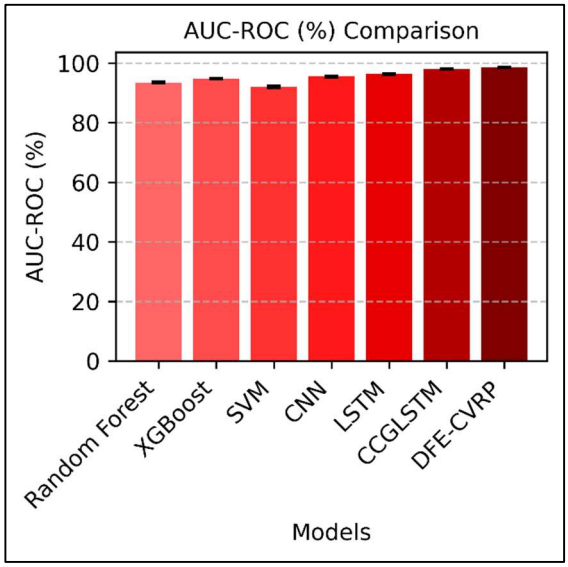

Figure 7: AUC-ROC Comparison Graph

5.5 AUC-ROC Analysis

The AUC-ROC (Area Under the Receiver Operating Characteristic Curve) serves as the vital measure to assess model discriminative performance regarding positive and negative CVD risk cases. A model with elevated AUC-ROC metrics demonstrates superior class differentiating capabilities therefore it serves as a vital performance assessment tool for medical diagnosis. Figure 7 shows different models' AUC-ROC comparative values. The AUC-ROC value of 98.6% for the DFE-CVRP model outperformed the base paper model's result of 98.0%. The dynamic ensemble learning strategy improves prediction accuracy because it enhances feature weighting and prediction calibration.

5.6 Discussion

The selection of optimum models for specific features that run in real-time boosts accuracy levels from 97.5% (CCGLSTM) to 98.2% (DFE-CVRP). The model demonstrates enhanced reliability in healthcare domain because it shows precision at 97.8% and recall at 98.4% thus minimizing false negative and false positive results.

The main factor behind accuracy improvement stems from cGANs data enhancement which improves minority class visibility while reducing model prejudice. The model's performance achieves optimal convergence through Successive Halving and Population-Based Training (PBT) methods while maintaining minimal computational expense. The ensemble learning along with feature-specific expert models in DFE-CVRP provides greater real-time adaptability than Random Forest and XGBoost models because these traditional machine learning methods lack dynamic features. Severe pattern detection capabilities of CNN and LSTM architectures are weakened by their limited ability to adjust features automatically in structured medical datasets. The model achieved AUC-ROC scores of 98.6% by using DFE-CVRP which represents an improvement over 98.0% with CCGLSTM thus demonstrating superior risk classification ability for patient care decision support and intervention planning.

6. CONCLUSION

The research proposed DFE-CVRP which serves as a deep learning framework designed to cope with major issues affecting standard CVD risk

models intended for diabetic patients. DFE-CVRP differentiates from standard models because it implements dynamic expert model selection of EfficientNet-based models augmented by cGAN features and optimized by Successive Halving and PBT methods.

A dynamic ensemble learning method stands as the main scientific achievement of this research which applies real-time validation accuracy to determine the network selection for feature-specialization. This adaptive model selection method leads to improved resilience and flexibility as well as interpretability which differentiates it from typical fixed model combinations. Data augmentation with cGAN enhances class imbalance management by creating diverse training sets which allows the model to build better generalization abilities for different patients.

The experimental findings established that DFE-CVRP described framework surpassed standard machine learning technologies Random Forest, XGBoost and SVM along with CNN, LSTM and CCGLSTM baseline deep learning models. The predictive capabilities of DFE-CVRP proved solid as shown through its 98.2% accuracy along with 97.8% precision and 98.4% recall and 98.1% F1-score and 98.6% AUC-ROC value.

Limitations: The current validation process of the framework operates on structured clinical data yet requires potential modifications before applying it to datasets with various demographic and clinical patterns. The data diversity improvement from cGANs does not effectively represent actual patient variations which are complex or appear rarely.

The future research will use DFE-CVRP by adding longitudinal health data, studying population generalization, and implementing federated learning to protect patient privacy across multiple centers.

REFERENCES:

[1] Timmis, A., Vardas, P., Townsend, N., Torbica, A., Katus, H., De Smedt, D., Gale, C. P., Maggioni, A. P., Petersen, S. E., Huculeci, R., Kazakiewicz, D. European Society of Cardiology: Cardiovascular Disease Statistics 2021. European Heart Journal, 2022, 43(8), 716-799.

[2] Navar, A. M., Fine, L. J., Ambrosius, W. T., Brown, A., Douglas, P. S., Johnson, K., Khera, A. V., Lloyd-Jones, D., Michos, E. D., Mujahid, M., Muñoz, D. Earlier Treatment in Adults with High Lifetime Risk of Cardiovascular Diseases: What Prevention Trials Are Feasible and Could Change Clinical Practice? Report of a National Heart, Lung, and Blood Institute (NHLBI) Workshop. American Journal of Preventive Cardiology, 2022, 12, 100430.

[3] Shishehbori, F., Awan, Z. Enhancing Cardiovascular Disease Risk Prediction with Machine Learning Models. arXiv preprint arXiv:2401.17328, 2024, Jan 29.

[4] Zhang, L., Lin, J., Liu, B., Zhang, Z., Yan, X., Wei, M. A Review on Deep Learning Applications in Prognostics and Health Management. IEEE Access, 2019, 7, 162415-162438.

[5] Kasartzian, D. I., Tsiampalis, T. Transforming Cardiovascular Risk Prediction: A Review of Machine Learning and Artificial Intelligence Innovations. Life, 2025, 15(1), 94.

[6] Thirunavukkarasu, J., Chinnasamy, A. Enhancing the Preciseness of Prediction in Heart Disease Diagnosis by Utilizing Machine Learning. Proceedings of the 2024 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), 2024, May 9, 1-9. IEEE.

[7] Moshawrab, M., Adda, M., Bouzouane, A., Ibrahim, H., Raad, A. Reviewing Multimodal Machine Learning and Its Use in Cardiovascular Diseases Detection. Electronics, 2023, 12(7), 1558.

[8] Abdullahi, M., Oyelade, O. N., Kana, F. A., Bagiwa, M. A., Abdullahi, F. B., Junaidu, S. B., Iliyasu, I., Ore-ofe, A., Chiroma, H. A Systematic Literature Review of Visual Feature Learning: Deep Learning Techniques, Applications, Challenges and Future Directions. Multimedia Tools and Applications, 2024, Jul 20, 1-58.

[9] Nithya, R., Kokilavani, T., Beena, T. L. Balancing Cerebrovascular Disease Data with Integrated Ensemble Learning and SVM-SMOTE. Network Modeling Analysis in Health Informatics and Bioinformatics, 2024, 13(1), 12.

[10] C Selvarathi, S. Varadhaganapathy. Deep Learning Based Cardiovascular Disease Risk Factor Prediction Among Type 2 Diabetes Mellitus Patients. Information Technology and Control, 2023, 52.0(1.0), 215-227.

[11] M. Almatari, Belal Abuhaija, Aladeen Alloubani, Firas Haddad, Ghaith M. Jaradat, Yousef Qawqzeh, Mutasem K. Alsmadi, Fahad AlGhamdi, Jehad Saad Alqurni, Lena Alodat, L

Dong. Cardiovascular disease risk factors prediction using deep learning convolutional neural networks. International Journal of Power Electronics and Drive Systems, 2024, 14.0(4.0), 4471-4471.

[12] Jingcheng Hu, Guangyu Hao, Jialiang Xu, Ximing Wang, Meng Chen. Deep learning-based coronary artery calcium score to predict coronary artery disease in type 2 diabetes mellitus. Heliyon, 2024, 10.0.

[13] S P Sachin Raj, Neelima Bayappu. Multimodal Deep Learning in Medical Diagnostics: A Comprehensive Exploration of Cardiovascular Risk Prediction. nan, 2024, 78-94.

[14] Shatabdi Das, Riaz Rahman, Ashis Talukder. Determinants of Developing Cardiovascular Disease Risk with Emphasis on Type-2 Diabetes and Predictive Modeling Utilizing Machine Learning Algorithms. nan, 2024.

[15] Mokammel Hossain Tito, Md Arifuzzaman, Alifa Nasrin, Shahzad Khan, M. Asaduzzaman, Muhammad Shahzad Chohan, Ali Nabil Al-Duais. Deep Learning for Prediction of Cardiovascular Disease. nan, 2024, 599-603.

[16] Arief Purnama Muharram, Fahmi Sajid. Supervised Machine Learning Approach for Predicting Cardiovascular Complications Risk in Patients with Diabetes Mellitus. nan, 2024.

[17] Ooi Ting Kee, Harmiza Harun, Norlaila Mustafa, Nor Azian Abdul Murad, Siok Fong Chin, Rosmina Jaafar, H H Che Hassan, M. Zubir Suboh, Noraidatulakma Abdullah. Cardiovascular Disease Risk in Type 2 Diabetes Mellitus Prediction Model Development Using Machine Learning. nan, 2023, 61-65.

[18] S.V. Evangelin Sonia, R. Nedunchezhian, M. Rajalakshmi. A multi-modal integrated deep neural networks for the prediction of cardiovascular disease in type-2 diabetic males. Automatika, 2023.

[19] Joon Yong Lee, Yera Choi, Taehoon Ko, Kang Hyuck Lee, Juyoung Shin, Hun-Sung Kim. Prediction of Cardiovascular Complication in Patients with Newly Diagnosed Type 2 Diabetes Using an XGBoost/GRU-ODE-Bayes-Based Machine-Learning Algorithm. Endocrinology and Metabolism, 2023.

[20] Jingzhi Yu, Xiaoyun Yang, Yu Deng, Amy E. Krefman, Lindsay R. Pool, Lihui Zhao, Xinlei Mi, Hongyan Ning, John T. Wilkins, Donald M. Lloyd-Jones, Lucia C Petito, Norrina B. Allen. Incorporating longitudinal history of risk factors into atherosclerotic cardiovascular disease risk prediction using deep learning. Dental science reports, 2024, 14.0.

[21] T. Abegaz, A. Baljoon, Oluwaseun Kilanko, Fatimah Sherbeny, Askal Ayalew Ali. Machine learning algorithms to predict major adverse cardiovascular events in patients with diabetes. nan, 2023, 164.0107289-107289.

[22] K.M. Hou, Chung-Hsing Chao. Predicting Cardiovascular Risk with Artificial Intelligence and Deep Learning. nan, 2024, 469-472.

[23] Jingjing Ren, Dong Liu, Guangpu Li, Jiayu Duan, Jiancheng Dong, Zhangsuo Liu. Prediction and Risk Stratification of Cardiovascular Disease in Diabetic Kidney Disease Patients. Frontiers in Cardiovascular Medicine, 2022, 9.0.

[24] Ooi Ting Kee, Harmiza Harun, Norlaila Mustafa, Nor Azian Abdul Murad, Siok Fong Chin, Rosmina Jaafar, Noraidatulakma Abdullah. Cardiovascular complications in a diabetes prediction model using machine learning: a systematic review. Cardiovascular Diabetology, 2023, 22.0(1.0).

[25] Mohamed, A. T., Santhoshkumar, S., Varadarajan, V. Intelligent Deep Learning-Based Predictive Model for Coronary Heart Disease and Chronic Kidney Disease on People with Diabetes Mellitus. Malaysian Journal of Computer Science, 2022, Mar 31, 88-101.

[26] Minhong Wang, Farah Francis, Holger Kunz, Xiang Zhang, Cheng Wan, Yun Liu, Paul Taylor, Sarah H. Wild, Honghan Wu. Artificial intelligence models for predicting cardiovascular diseases in people with type 2 diabetes: A systematic review. Intelligence-based medicine, 2022, 6.0100072-100072.

[27] García-Ordás, M. T., Bayón-Gutiérrez, M., Benavides, C., Aveleira-Mata, J., Benítez-Andrades, J. A. Heart Disease Risk Prediction Using Deep Learning Techniques with Feature Augmentation. Multimedia Tools and Applications, 2023, 82(20), 31759-31773.

[28] Ruilei Pang, Fang Fang, Yang Wu, Fangfang Li, Ruixin Dong, Tianzuo Zhang. Analysis and prediction of type-2 diabetes mellitus complicated with coronary heart disease based on LSTM. nan, 2023, 12724.0127240K-127240K.

[29] Abhilash Mayya, H. Solieman. Machine Learning System for Predicting Cardiovascular Disorders in Diabetic Patients. IzvestiÃ¢ vysÅ¡ih uÄ ebnyh zavedenij Rossii, 2022, 25.0(4.0), 116-122.

[30] Dongzhe Hong, Daniel Fort, Lizheng Shi, Eboni G. Price-Haywood. Electronic Medical Record Risk Modeling of Cardiovascular Outcomes Among Patients with Type 2 Diabetes. Diabetes Therapy, 2021, 12.0(7.0), 2007-2017.

[31] Rui Fan, Ning Zhang, Longyan Yang, Jing Ke, Dong Zhao, Qinghua Cui. AI-based prediction for the risk of coronary heart disease among patients with type 2 diabetes mellitus. Scientific Reports, 2020, 10.0(1.0), 14457-14457.

[32] Madhuri Panwar, Arvind Gautam, Rashi Dutt, Amit Acharyya. CardioNet: Deep Learning Framework for Prediction of CVD Risk Factors. nan, 2020, 1-5.

[33] Gabriela Roman, Anca Pantea Stoian. Cardiovascular Risk/Disease in Type 2 Diabetes Mellitus. nan, 2021.

[34] Ahmad, A., Lim, L. L., Morieri, M. L., Tam, C. H., Cheng, F., Chikowore, T., Dudenhöffer-Pfeifer, M., Fitipaldi, H., Huang, C., Kanbour, S., Sarkar, S. Precision Prognostics for Cardiovascular Disease in Type 2 Diabetes: A Systematic Review and Meta-Analysis. Communications Medicine, 2024, 4(1), 11.

[35] Ekramul Hossain, Shahadat Uddin, Arif Khan. Network analytics and machine learning for predictive risk modelling of cardiovascular disease in patients with type 2 diabetes. Expert Systems With Applications, 2021, 164.0113918.

[36] Pytlak, K. Personal Key Indicators of Heart Disease. Kaggle Dataset, 2022. Available at: https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease/data.