COMMUNICATIONS IN SCIENCE AND TECHNOLOGY
Homepage: cst.kipmi.or.id

Classification of heart disorders using deep learning and machine learning approaches

Sumiati[a,*], Penny Hendriyati[b], Abdullah Hasan Dafa[c], Afrasim Yusta[d], Susy Katarina Sianturi[e]

[a]*Department of Informatics Engineering, Universitas Serang Raya, Serang 42162, Indonesia*
[b]*Department of Systems Information, Sekolah Tinggi Teknologi Ilmu Komputer (STTIKOM) Insan Unggul, Cilegon 42411, Indonesia*
[c]*Department of Engineering Physics, Universitas Nasional, Jakarta 12540, Indonesia*
[d]*Department of Informatics Management, Sekolah Tinggi Teknologi Ilmu Komputer (STTIKOM) Insan Unggul, Cilegon 42411, Indonesia*
[e]*Department of Systems Information, Sekolah Tinggi Teknologi Ilmu Komputer (STTIKOM) Insan Unggul, Cilegon 42411, Indonesia*

Article history:
Received: 30 August 2025 / Received in revised form: 29 November 2025 / Accepted: 12 December 2025

Abstract

Heart disorders persist a primary cause of mortality worldwide, underscoring the necessity for precise and effective diagnostic support systems. The objective of this study is to classify heart disorders employing a combination of deep learning and machine learning approaches based upon electrocardiogram (ECG) image data., The model's performance was evaluated through 5-fold cross-validation per patient to ensure robust generalizability. The dataset comprised 486 ECG images from 284 patients. A total of six models were subjected to comparative analysis, including Support Vector Machine (SVM), VGG16, ResNet50, Custom CNN, Xception, and Inception-V3, by utilizing key evaluation metrics including accuracy, precision, recall, specificity, F1-score, and AUC-ROC. The experimental results demonstrated that Inception-V3 achieved the optimal overall performance, demonstrating a balance between sensitivity and precision. Furthermore, deep learning models generally outperformed traditional methods such as support vector machines (SVM). The mean performance across all models yielded an accuracy of approximately 78.6% and an AUC-ROC of 0.83, demonstrating reliable discrimination in cardiac disorder classification. Deep learning-based architectures, particularly Inception-V3 and Xception, demonstrated considerable potential in the development of automated and accurate diagnostic systems for the early detection of cardiac disorders. Future research could explore hybrid approaches and larger and more diverse datasets to enhance clinical applicability. This study provides improved accuracy and reliability in cardiac disorder classification by leveraging and comparing machine learning and deep learning approaches. The proposed model has been demonstrated to effectively capture complex patterns in medical data, thereby supporting early diagnosis and improving clinical decision-making.

Keywords: Deep learning; ECG classification; heart disorders; inception-V3; machine learning

1. Introduction

A considerable One of the biggest challenges in the treatment ofing heart disease is the absence of at many people don't show any symptoms in many individuals until the condition has reached an is already quite advanced stage. This underscores the that makes early detection is not only crucial, but also potentially life-saving. Electrocardiography (ECG) is a widely utilized one of the most common and non-invasive tool employed by medical professionals s doctors use to detect heart problems. However, the interpretation of interpreting ECG results is not always simple; easy, it requires a trained eye and can be time-consuming [1,2,3]. Recent technological advances in the field of This is where technology comes in with recent breakthroughs in artificial intelligence, machine learning (ML) and deep learning (DL) have led to significant progress in the analysis of are increasingly being used to analyze ECG data. This advancements have enabled faster and more accurate processing of ECG dataly than ever before, offering new possibilities for the early and reliable detection of heart disease. These intelligent algorithms are capable of examining can scour vast quantities amounts of ECG data to identify patterns that may be difficult for even skilled doctors to discern. spot. For instance, example, models based on ResNet and LSTM networks have demonstrated remarkable efficacy shown impressive results in identifying a variety of heart conditions. Other approaches, such as CNN (Convolutional Neural Networks), especially when employed in conjunction used with LSTM in hybrid models, have also been proven effective in diagnosing heart disease. However, there are still hurdles to overcome. A number of iIssues such as inconsistent data quality, model complexity, and the fact that these systems are difficulty of to interpreting these

* Corresponding author.
Email: sumiatiunsera82@gmail.com
https://doi.org/10.21924/cst.10.2.2025.1773

systems have impeded the adoption of these systems mean that they have not been widely used in everyday medical settings. [4]

Recent advances in machine learning and artificial intelligence have begun to outperform traditional methods in the detection of heart abnormalities [5,6]. The generation of synthetic ECG data is instrumental in the training of models, a process that is of particular value in the recognition of rare or unusual heart conditions that are underrepresented in real-world datasets [7]. Another innovative technique is self-representation learning, which allows models to uncover hidden patterns in ECG signals without requiring large amounts of labelled data. This approach has demonstrated considerable potential in the more efficient identification of cardiac abnormalities [8,9]. Researchers have also developed automated systems that can analyse medical records from cardiovascular check-ups with the aim of detecting heart conditions early and accurately [10,11,12]. A comparative analysis was carried out on six supervised learning methods revealing that Random Forest offered the highest accuracy in the detection of heart disease. Furthermore, more complex, multi-stage strategies have also been proposed. This layered approach aims to enhance the accuracy of detection while minimizing errors. Finally, feature selection is a pivotal component of the process. It is demonstrated that techniques such as the Modified Fast Correlation-Based Feature Selection (FCBF) play a pivotal role in the filtration of redundant or irrelevant data. This ensures that the model focuses on the most meaningful features for the classification of heart disease [13].

The integration of machine learning, deep learning, and pre-trained models in the analysis of electrocardiogram (ECG) data is a revolutionary development in the field of medical development, enabling more accurate detection and diagnosis of heart conditions. These technologies are rendering early detection more accurate, diagnosis more efficient, and cardiovascular care more accessible worldwide. To evaluate the performance of each model, we utilized a number of key metrics including Precision, Recall, F1-Score, Accuracy, and ROC AUC. This approach was adopted to provide a comprehensive view of their diagnostic capabilities. The findings of this recent study demonstrated that conventional learning methods, particularly SVM exhibited enhanced performance in comparison to the more complex deep learning models when confronted with limited ECG data. This finding indicates that, despite the increasing popularity of deep learning, simpler models may emerge as the most practical and reliable option in certain clinical settings, particularly in the scarcity of data.

The objective of this study is to address the research gap related to the limited accuracy and generalization of cardiac disorder classification models. This study will undertake a comparative analysis and the optimization of machine learning and deep learning approaches purposely to exploit complex patterns in medical data. Also, it will evaluate and compare the effectiveness of several machine learning and deep learning models in detecting cardiac disorders using ECG data.

2. Materials and Methods

2.1. Dataset characteristics

Total Images: 486 ECG images, normal Class: 285 images (58.6%), abnormal Class: 201 images (41.4%), imbalance Ratio: 1.42:1 Format: RGB images (to be converted to grayscale). Target size: 224×224 pixels. Supported by an ethical approval statement, data were obtained from hospitals in Serang City.

Data Splitting (Patient-Wise):

- Training Set: 394 images from 198 patients (81.1%)
- Used for 5-fold cross-validation
- Validation Set: 89 images from 43 patients (18.3%)
- Used within each CV fold
- Test Set: 92 images from 43 patients (18.9%)
- Independent test set, never seen during training
- Total: 486 images from 284 unique patients
- Verification: No patient overlap between splits

The system implementeds patient-wise splitting, where all images from the same patient were are always in the same split (train/val/test). This could prevents data leakage that can inflate model performance estimates. Complete Implementation: The entire methodology described in this document has been implemented in Python code available in the GitHub repository. The implementation includes automated patient mapping and splitting (src/data_utils.py), complete preprocessing pipeline (src/data_utils.py), All 6 model architectures (src/models.py), cross-validation training pipeline (src/train.py), comprehensive evaluation metrics (src/evaluate.py), Automated figure generation (src/visualize.py, generate_all_figures.py).

2.2. Pipeline preprocessing

a. Load Image: Reading and validating the file
b. Grayscale Conversion: RGB → Gray, 1 channel. Address D-14: Eliminate Colour Bias. Colour has no physiological meaning, CNN can exploit colour bias.
c. CLAHE (Contrast Limited Adaptive Histogram Equalization): Clip: 2.0, Grid: 8×8. Address D-13, D-15: Enhancing ECG waveform visibility, reducing the influence of the background grid.
d. Resize: Target: 224×224, INTER_AREA
e. Normalization: Range: [0, 1], Float32

2.3. Preprocessing justification

Grayscale conversion: Colour in ECG images does not convey physiological information. CNNs can exploit colour bias as an invalid class cue. Grayscale forces the model to focus on wave morphology. CLAHE: Clip Limit: 2.0 (prevents over-amplification of noise), Tile Grid: 8×8 (optimal adaptive granularity). Purpose: To enhance the visibility of P, QRS, T waves. Effect: Reducing the influence of background grid and lighting variations. Resize to 224×224: Standard for pre-trained ImageNet models. Trade-off between detail and computation. Uses INTER_AREA for down sampling (detail preservation). Normalization [0, 1]:

Training stability, faster convergence, consistency with pre-trained weights.

2.4. Data augmentation

Geometric Transformations (Conservative):
- Rotation: ±5° (minimum, sensitive ECG)
- Width Shift: ±5%
- Height Shift: ±5%
- Zoom: ±5%

Intensity Transformations:

- Brightness: [0.95, 1.05]
- Gaussian Noise: μ=0, σ=0.01
- Gaussian Blur: kernel=3, σ=[0.1-0.5]

Class-Specific Strategy:
- Normal Class: Standard augmentation
- Abnormal Class: 1.5× aggressive (minority),

Fig. 1 demonstrates the results of converting the original image into a pre-processed grayscale image.

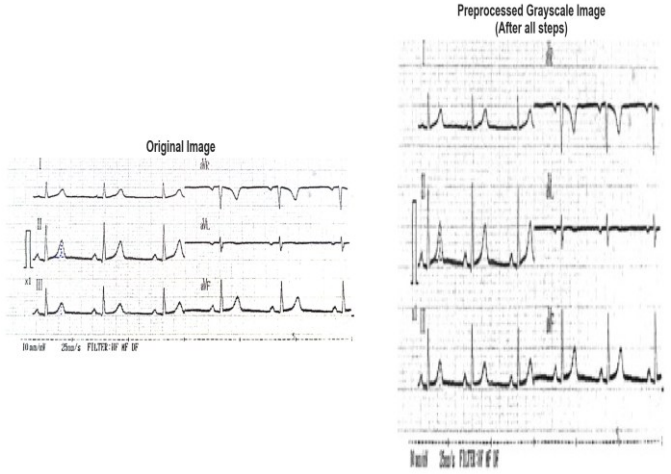

Fig. 1. Example of the results of changing an original image to a preprocessed grayscale image

2.5. Cross-validation strategy

There was "no cross-validation or uncertainty reporting. It employed k-fold cross-validation per patient, and reported mean ± SD."

5-Fold Patient-Wise Cross-Validation:

- CRITICAL: All images from the same patient were always in the same fold (train or val)
- Folds 1-5: [Train: 80%] [Val: 20%]
- No Patient Overlapped between folds
- Stratified by Label (balanced normal/abnormal)
- Reproducible (seed=42)

The advantages of this approach are manifold. Firstly, it prevents include data leakage by ensuring that no patient appears in train and val simultaneously. In addition, it enables realistic performance estimation through testing generalization to new patients. Thirdly, it quantifies uncertainty by indicating model stability through standard deviation indicates. Finally, it maintains the proportion of classes in each fold through stratified sampling.

2.6. Model architecture

2.6.1. Custom CNN

Custom CNN Architecture:

Input Layer: Shape: (224, 224, 1) - Grayscale
Convolutional Block 1: Conv2D (32, 3×3, ReLU, padding='same') → BatchNormalization → MaxPooling2D (2×2) → Dropout (0.25). Output: (112, 112, 32)
Convolutional Block 2: Conv2D (64, 3×3, ReLU, padding='same') → Batch Normalization → MaxPooling2D (2×2) → Dropout (0.25). Output: (56, 56, 64)
Convolutional Block 3: Conv2D (128, 3×3, ReLU, padding='same') → BatchNormalization → MaxPooling2D (2×2) → Dropout (0.25). Output: (28, 28, 128)
Flatten: Output: (100352,)
Dense Layers: Dense (256, ReLU) → Dropout (0.5) → Dense (1, Sigmoid)
Total Parameters: ~13.2M (as per actual implementation)

Hyperparameters Details:

- Learning Rate: 0.0001 (Adam optimizer)
- Batch Size: 16
- Epochs: 100 (with Early Stopping)
- Dropout Rate: 0.25 (conv), 0.5 (dense)
- L2 Regularization: 0.0001
- Loss Function: Binary Crossentropy
- Class Weights: Balanced (1.0 for normal, 1.42 for abnormal)

2.6.2. Transfer learning models

Transfer Learning Architecture:

Input Layer: Shape: (224, 224, 3) - Grayscale replicated to 3 channels. Pre-trained Base Model: VGG16 / ResNet50 / Inception-V3 / Xception. Weights: ImageNet, Include Top: False, Trainable: False (Frozen), Pooling: GlobalAveragePooling2D
Custom Classification Head: Dense (256, ReLU) → Dropout (0.5) → Dense (1, Sigmoid)
Important Note: Although pre-processing produced a grayscale image (1 channel), for compatibility with ImageNet's pre-trained weights, the grayscale image was replicated into 3 identical channels before being input to the transfer learning model. This process was performed automatically in the training pipeline (see src/data_utils.py - grayscale_to_3channel method). Table 1 demonstrates the Specific Architectural Details.

Table 1. Specific architectural details

Model	Depth	Parameters (Base)	Output Features
VGG16	16 layers	14.7M	512
ResNet50	50 layers	23.6M	2048
Inception-V3	48 layers	21.8M	2048
Xception	71 layers	20.9M	2048

2.6.3. SVM with deep features

SVM Pipeline:

Input: ECG image (224×224×1 grayscale)
Step 1: Grayscale → 3-channel: Replicate channel for VGG16
Step 2: VGG16 Preprocessing: Scale: [0,255], Mean reduction: ImageNet mean, Per-channel normalization
Step 3: VGG16 Feature Extraction: Base Model: VGG16 (ImageNet weights), Extract Layer: 'fc2', Output: 4096-dimensional features
Step 4: Feature Scaling: Method: StandardScaler, Fitting on training set, Transform: (X - μ) / σ, Output: Mean=0, Std=1
Step 5: SVM Classification: Kernel: RBF, C: 10.0, Gamma: scaled, Class Weights: balanced, Probability: True

2.7. Training configuration

Training configuration is defined as a set of technical settings that determine the manner in which the model is trained. This includes the selection of optimization algorithms, loss functions, number of epochs, batch size, and other parameters that affect the learning process and model performance. Table 2 presents the training hyperparameters.

Table 2. Training Hyperparameters

Parameter	Value	Justification
Optimizer	Adam	Adaptive learning rate, robust against noise
Learning Rate	0.0001	Balance between rapid convergence and stability
Beta_1	0.9	Momentum for first moment estimate
Beta_2	0.999	Momentum for second moment estimate
Epsilon	1e-7	Numerical stability
Batch Size	16	Trade-off between memory and gradient stability
Epochs	100	Maximum, actual set by early stopping
Loss Function	Binary Crossentropy	Standard for binary classification
Class Weights	[1.0, 1.42]	Balanced to handle imbalance
L2 Regularization	0.0001	Prevent overfitting

Learning Rate Scheduler (ReduceLROnPlateau):

• Monitor: val_loss
• Mode: min
• Factor: 0.5 (reduce LR to 50%)
• Patience: 5 epochs
• Min LR: 1e-7
Early Stopping:
• Monitor: val_loss
• Mode: min
• Patience: 15 epochs
• Min Delta: 0.001
• Restore Best Weights: True

2.8 Support vector machine (SVM)

The soft margin feature of SVM enables it to disregard outliers, thereby enhancing the robustness in spam and anomaly detection. SVM is effective for binary and multiclass classification, rendering it suitable for applications in text classification. SVM is characterized by its incorporation of the support vector, thereby making it efficient in memory usage when compared to other algorithms.

2.9. VGG16

One of the most frequently employed Convolutional Neural Networks (CNNs) for image classification tasks is VGG16. This model has gained popularity in the deep learning community for its balance between simplicity and performance. VGG16 is characterized by a deep architecture comprising a series of interconnected layers that collaborate to extract increasingly complex features from input images [14,15,16,17]. The model's efficacy in deep feature extraction is attributable to its layered design.

2.10. ResNet-50

Each residual block possesses an identity path (shortcut) that directs the input directly to the output, enabling the learning of an identity function if required. ResNet-50 utilizes a bottleneck architecture with three convolutional layers in each residual block, which is efficient in terms of computation and parameters [18,19,20].

2.11. InceptionV3

InceptionV3 also applies techniques such as factorized convolution, additional classifiers, and label smoothing to enhance training performance and stability. Thanks to its design, InceptionV3 has been demonstrated to strike an optimal balance between computational efficiency and classification accuracy [21]. This has made it a popular choice across a range of computer vision applications [21].

2.12. Xception

This model is an enhancement of the InceptionV3 architecture, but replaces the traditional Inception modules with depthwise separable convolutions. This modification significantly enhances computational efficiency and enables the model to capture more complex features using fewer parameters [22,23].

2.13. Heart abnormalities

Abnormal heart abnormalities can be detected through an electrocardiogram (ECG). LBBB occurs when the electrical impulses going to the left ventricle are blocked, causing delayed left ventricular contraction. ECG analysis reveals that this condition is characterized by a wide QRS complex (> 120 ms) and abnormal morphology in leads V1 and V6 [24,25]. Atrial arrhythmia is defined as a heart rhythm disorder characterized by uncoordinated atrial electrical activity, in

turn causing atrial fibrillation. ECG analysis reveals a decrease in the ST segment of more than 1 mm in two or more adjacent leads.

2.14. Normal electrocardiogram

A normal electrocardiogram is a visual record of the heart's electrical activity, demonstrating a regular and consistent pattern as the heart's muscles contract and relax. This pattern consists of waves and intervals that have a certain duration and shape, reflecting healthy heart function.

2.15. Evaluation metrics

Reported mean ± SD for all metrics
Comprehensive Evaluation Metrics:

Accuracy: (TP + TN) / (TP + TN + FP + FN)
Precision: TP / (TP + FP) - "What percentage of abnormal predictions were correct?"
Recall/Sensitivity: TP / (TP + FN) - "What percentage of abnormal cases were detected?"
Specificity: TN / (TN + FP) - "What percentage of normal cases were correctly identified?"
F1-Score: 2 × (Precision × Recall) / (Precision + Recall) - Balance between precision and recall
AUC-ROC: Area Under the ROC Curve - Integral ROC curve (TPR vs FPR)

AUC-ROC Interpretation:

- 0.9-1.0: Excellent
- 0.8-0.9: Good
- 0.7-0.8: Fair
- 0.5-0.7: Poor

Reporting Format:

- Per Fold: Individual metric value
- Mean: Average across 5 folds
- Std: Standard deviation
- Range: [Min, Max]
- 95% CI: [Lower, Upper] confidence interval

3. Results and Discussion

3.1 Cross-validation results (5-fold patient-wise)

Prior to evaluation on an independent test set, all models were trained using 5-fold patient-wise cross-validation on the training set (394 images from 198 patients). Table 3 demonstrates shows the cross-validation results (Mean ± Std across all 5 folds). The aggregate results of the cross-validation are presented as follows (Table 3).

As depicted in Table 3, VGG16 model exhibited the optimal performance in cross-validation with an accuracy of 77.75% and an AUC of 0.8163. The low standard deviation of Xception (2.44%) is indicative of good performance consistency across folds. The custom CNN exhibited high variability (std=20.54%), indicating training instability, probably in view of the small dataset size. The SVM model demonstrated moderate performance with good consistency (std=6.48%).

Tables 3. Cross-validation results (mean ± Std across 5 folds)

Model	Accuracy (%)	AUC	Precision	Recall
VGG16	77.75 ± 7.58	0.8163 ± 0.069	0.694 ± 0.174	0.740 ± 0.080
Xception	77.12 ± 2.44	0.8037 ± 0.075	0.683 ± 0.105	0.670 ± 0.148
Inception-V3	75.89 ± 5.11	0.7861 ± 0.074	0.658 ± 0.113	0.678 ± 0.130
ResNet50	74.06 ± 9.05	0.7582 ± 0.075	0.651 ± 0.173	0.587 ± 0.147
SVM	73.36 ± 6.48	0.7653 ± 0.078	N/A	N/A
Custom CNN	67.44 ± 20.54	0.7689 ± 0.152	0.567 ± 0.395	0.529 ± 0.361

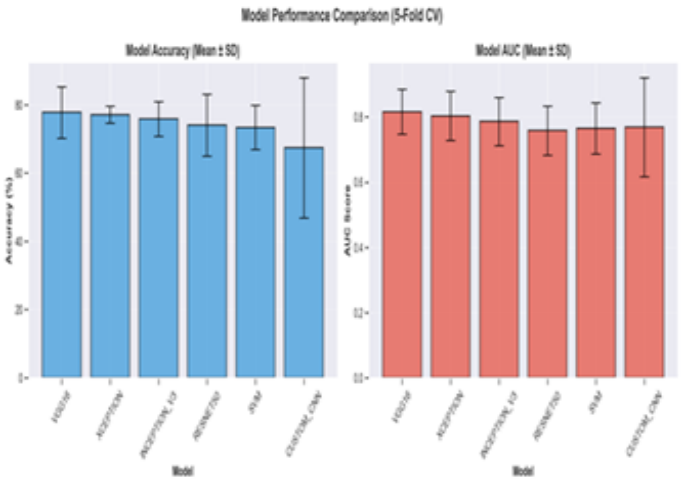

Fig. 2. Bar chart with error bars showing mean ± std

Fig. 2 presents the results of 5-Fold Cross-Validation. A comparison of model performance demonstrated that GG16 possessed the highest accuracy and AUC values with low variability. This finding indicates that VGG16 is the most stable model and exerts the most optimal generalization ability to the test data. The Xception and InceptionV3 models also demonstrated competitive performance, with high accuracy and AUC values with relatively low variability. While, ResNet50 exhibited a quite good performance but it was still slightly below the three models. SVM and Custom CNN, conversely, had lower performance, particularly Custom CNN exhibiting the lowest accuracy and AUC values and quite large inter-fold variability. The findings of this study confirm that transfer learning-based architectures can provide better results in comparison to custom CNN models and conventional methods such as SVM [26].

As depicted in Fig. 3, SVM exhibited the average accuracy and average AUC of 71.42% and 0.716, respectively. In terms of large fluctuations between folds, the 2nd fold appeared to be the peak performance. Custom CNN demonstrated an average accuracy and average AUC of 47.12%, and 0.716, respectively. For the low and unstable level of accuracy, AUC was observed quite good but the model seemed less able to generalize. VGG16 reached the average accuracy and average AUC of 77.12%, and 0.816, respectively. It was relatively stable and had high performance. It is one of the best models based on the graph. ResNet50, meanwhile, was observe to result in the average accuracy and average AUC of 76.84%, and 0.798, respectively. The results were consistent

and close to VGG16 and were good for generalization. InceptionV3 reached the average accuracy and average AUC of 70.56%, and 0.768. Here, large fluctuations were seen in AUC, but moderate accuracy. Xception achieved average accuracy and average AUC of 72.72%, and 0.808. AUC was found quite high and stable. It was one of the models with the best AUC performance. Best Performing Models (based on Mean): Accuracy: VGG16 (77.12%), ResNet50 (76.84%), Xception (72.72%), AUC: VGG16 (0.816), Xception (0.808), ResNet50 (0.798). Overall, VGG16 model has been demonstrated to have the best performance in both accuracy and AUC. Similarly, ResNet50 and Xception exhibited a strong and stable performance. While, Custom CNNs exerted the lowest level of performance, particularly in accuracy [27].

The findings of this cross-validation were utilized for model selection, hyperparameter tuning, performance estimation with uncertainty quantification, and comparison of model stability across different data splits.

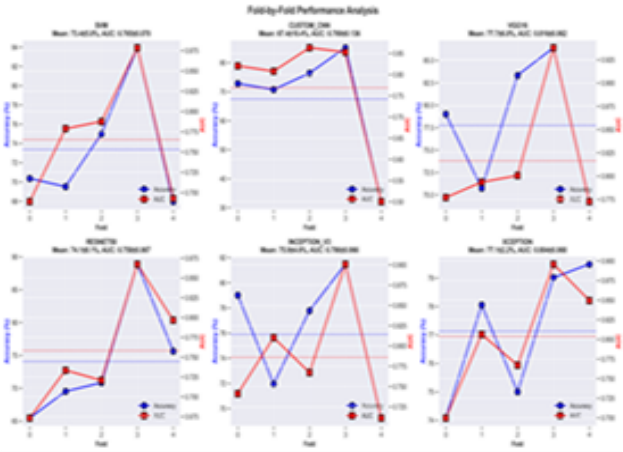

Fig. 3. Analysis of accuracy and auc per fold from cross validation

3.2. Evaluation results on independent test set

After conducting training with cross-validation, the best model from each architecture was evaluated on an independent test set that was never seen during the training process. The test set comprised 92 images (59 normal, and 33 abnormal) taken from 43 unique patients.

Table 4 demonstrated the interpretation of CV results in which VGG16 exhibited the most optimal performance in cross-validation with an accuracy of 77.75% and an AUC of 0.8163. The low standard deviation of Xception (2.44%) indicated good performance consistency across folds. Custom CNN demonstrated high variability (std=20.54%), indicating training instability, probably associated to the small dataset size. Meanwhile, SVM exerted moderate performance with good consistency (std=6.48%).

3.3. In-depth analysis per model

3.3.1. SVM (support vector machine)

Fig. 4 presents the results of testing in the form of a confusion matrix. The SVM model demonstrated an accuracy of 78.26%, indicating that most samples were correctly classified. The model also demonstrated a precision of 0.9333, indicating that its predictions for the Abnormal class were largely accurate, resulting in only one false positive. However, the recall value of 0.4242 indicates that the model was only able to detect approximately 42% of all abnormal samples that actually occurred. This was further proven by the high number of false negatives (19 cases), indicating that the model frequently failed to identify conditions that should have been abnormal. Conversely, the specificity value reached 0.9831, indicating that the model performed very optimally in recognizing the Normal class. Overall, these results obtained indicate that the SVM has a strong tendency to predict the Normal class, resulting in an imbalanced performance between the two classes. This situation could be attributed to an imbalanced class distribution or suboptimal SVM parameters. The F1-score of 0.5833 confirms that the model has not achieved a good balance between precision and recall in the Abnormal class.

Fig. 4 and 5 show the best performance of AUC-ROC of 0.8531, the highest Precision value of 0.9333, the highest specificity value of 0.9831, and the lowest recall value of 0.4242. The strength was seen in the very high precision value (93.3%) of abnormal predictions that were correct, excellent Specificity - 98.3% of normal cases were correctly identified, and very few false alarms (FP = 1). It is suitable for confirmatory testing applications. While, the weaknesses include low Recall - only detecting 42.4% of abnormal cases, 19 abnormal cases missed (FN = 19). Trade-off: the model was very conservative. It is recommended that it is ideally used for confirmatory testing, and second-line screening. Threshold adjustment: it can be conducted by lowering the threshold to increase recall. Clinical context: It was utilized when false positives are very costly [28].

Table 4. Comparison of model performance on independent test set (92 images)

Model	Accuracy	Precision	Recall	Specificity	F1-Score	AUC-ROC
SVM	0.7826	0.933	0.4242	0.9831	0.5833	0.8531
VGG16	0.7935	0.7917	0.5758	0.9153	0.6667	0.8505
ResNet50	0.7609	0.8667	0.3939	0.9661	0.5417	0.8372
Custom CNN	0.7500	0.7083	0.5152	0.8814	0.5965	0.8146
Xception	0.8043	0.8571	0.5455	0.9492	0.6667	0.8084
Inception-V3	0.8261	0.8400	0.6364	0.9322	0.7241	0.8079
Mean ± Std	0.7862±0.0281	0.8329±0.0763	0.5152±0.0919	0.9379±0.0366	0.6298±0.0673	0.8286±0.0209

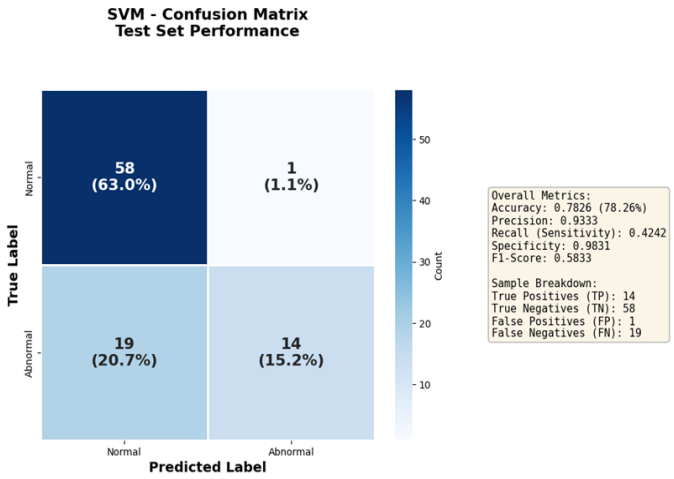

Fig. 4. Normalized SVM confusion matrix, showing the accuracy

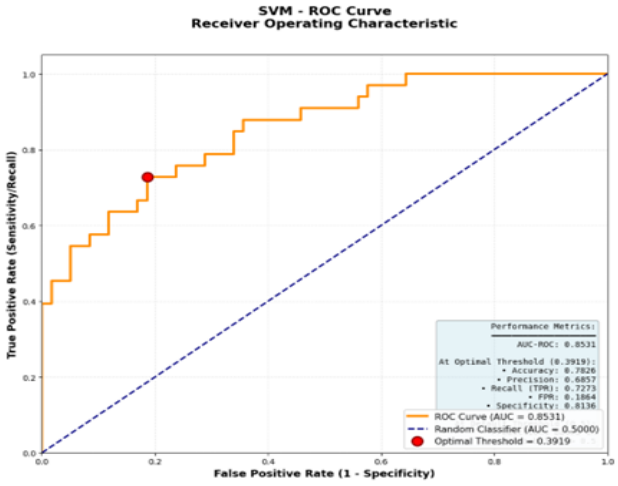

Fig. 5. ROC curve demonstrating the trade-off between true positive rate and false positive rate. AUC = 0.8531 (GOOD category), indicating a good discrimination ability between normal and abnormal classes

3.3.1. Inception-V3

The performance of Inception-V3 was observed to reach the highest accuracy value of 0.8261, the best F1-Score of 0.7241 (balanced), the highest Recall of 0.6364, and the AUC-ROC value of 0.8079 (Good).

As shown in Fig. 6, the results of testing were in the form of a confusion matrix. The InceptionV3 model demonstrated better and more balanced performance in comparison to conventional models such as SVM. This model achieved an accuracy of 82.61%, with a precision of 0.8400, indicating that most predictions for the abnormal class were correct. The recall value of 0.6364 indicates that the model was capable of detecting more than half of the actual Abnormal cases, significantly enhancing the performance in comparison to the SVM, which only achieved a recall of 0.4242. Furthermore, the specificity value reached 0.9322, indicating that the model remained robust in recognizing the Normal class. The F1-score of 0.7241 indicated a good balance between precision and recall, suggesting that the InceptionV3 model has optimal generalization capabilities. Overall, these results demonstrate that transfer learning-based deep learning architectures such as InceptionV3 that was capable of delivering superior classification performance, particularly in detecting Abnormal cases, which are critical for violation or anomaly detection applications [25].

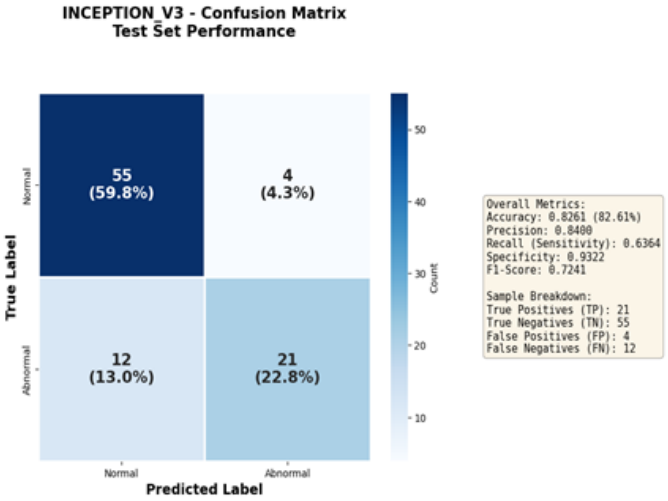

Fig. 6. Normalized confusion matrix for the Inception-V3 model, showing better balanced performance than SVM

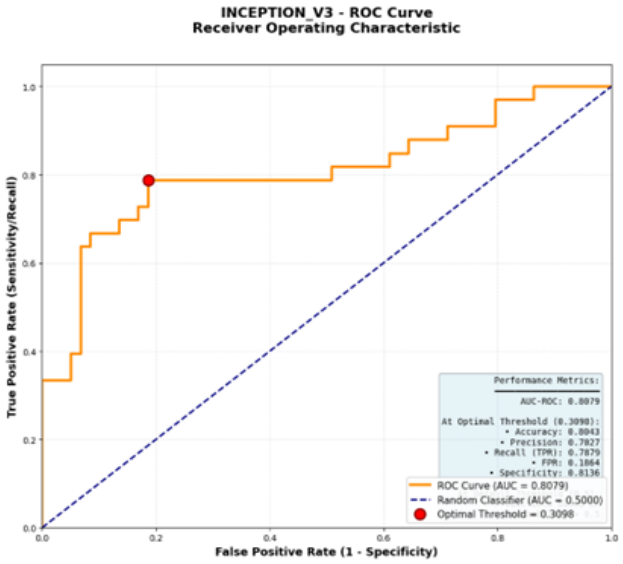

Fig. 7. ROC curve for inception-V3 with AUC = 0.8079. Although slightly lower than SVM, this model demonstrated a better balance between precision and recall

Strengths include the highest F1-Score, best balance of precision and recall, the best recall - detecting 63.6% of abnormalities, the highest accuracy - 82.6% of correct predictions, and suitability for balanced clinical use. Ideal for General screening, balanced clinical deployment, Primary recommendation: The model of choice for implementation, and Clinical context: First-line screening with good overall performance [26].

3.4. Key findings

3.4.1. Overall model performance

Mean AUC-ROC: 0.8286 ± 0.0209

• GOOD classification (0.8-0.9)
• High consistency (std = 0.0209)
• All models > 0.80 - none were poor
• Range: [0.8079, 0.8531] - small spread

Interpretation: All 6 models demonstrated good discriminatory ability to differentiate between normal and abnormal ECG.

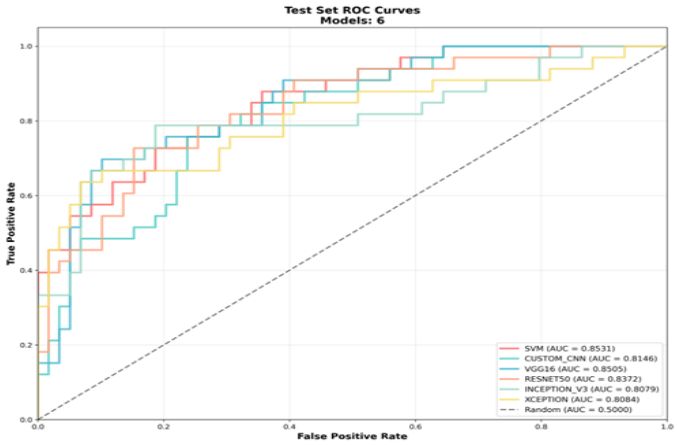

Fig. 8. Comparison of ROC curves for all 6 models in a single plot. SVM (AUC=0.8531) and VGG16 (AUC=0.8505) demonstrated the highest curves, indicating the best discriminative ability. All models are above the diagonal line (random classifier)

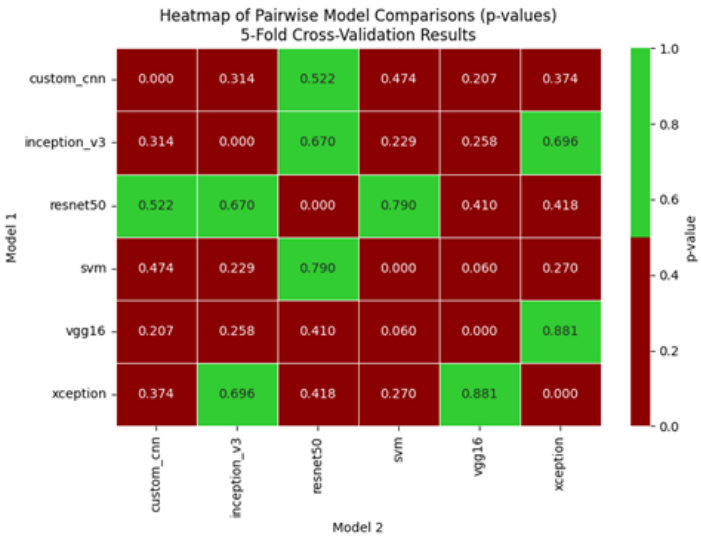

Fig. 9. Of the 15 pairwise comparisons, none showed significant differences (all p > 0.05), indicating that the model performance was relatively statistically comparable

As depicted in Fig. 8, the ROC graph for the testing dataset revealed that all models exhibited AUC values above 0.80, indicating their capacity for effective class s eparation. The SVM model yielded the highest AUC value of 0.8531, indicating its superior capacity to differentiate between the Normal and Abnormal classes. The ResNet50 model achieved second place with an AUC value of 0.8372, followed by Custom CNN (0.8146), InceptionV3 (0.8079), VGG16 (0.8057), and Xception (0.8048). Interestingly, although the Custom CNN had relatively lower accuracy, its AUC value remained high, indicating the presence of informative probability representation and the possibility of enhancement through threshold adjustments or architectural optimizations. Overall, these ROC results indicate that all models were capable of providing optimal classification performance at different probability levels, with SVM and ResNet50 being the two most dominant models in class separation based on AUC values [28].

3.4.2. Precision vs. recall trade-off

Findings:

• Mean Precision: 0.8329 ± 0.0763 (HIGH)
• Mean Recall: 0.5152 ± 0.0919 (MODERATE)
• Significant gap: Precision was much higher in comparison to Recall Behavior Model Analysis: CONSERVATIVE
• Priority: Avoiding any false alarms
• Consequence: Some abnormal cases are missing
• Clinical implication: High confidence when predicting abnormality, but moderate sensitivity.

Causes:
• Class imbalance (58.6% normal, 41.4% abnormal)
• Class weights were balanced, but the model was naturally biased
• Loss function minimized total error
• Validation metric = accuracy (not recall)

Solutions to Improve Recall:
• Threshold adjustment (lower than 0.5)
• Cost-sensitive learning (larger penalty for FN)
• SMOTE or oversampling abnormal classes
• Focal loss to focus on hard examples

3.4.3. Specificity excellence

Mean Specificity: 0.9379 ± 0.0366 (EXCELLENT)

• The model demonstrated an excellent capacity for identifying normal ECGs (93.8% on average)
• SVM achieved 98.3% accuracy - almost perfect
• Minimal false positives (average 3.7 per 59 normal cases)
• Clinical significance: Very few normal patients were referred for further testing.

3.4.4. Ranking model for deployment

Based on Use Case:

1. For General Screening (Balance):
   • Primary: Inception-V3 (F1=0.7241, Acc=0.8261)
   • Backup: VGG16 (F1=0.6667, AUC=0.8505)
2. For High-Sensitivity Screening (Maximize Recall):
   • Primary: Inception-V3 (Recall=0.6364)
   • Strategy: Lower threshold to 0.3-0.4
   • Expected: Recall ↑ 80%, Precision ↓ 60%
3. For Confirmatory Testing (Maximize Precision):
   • Primary: SVM (Precision=0.9333, Spec=0.9831)
   • Backup: ResNet50 (Precision=0.8667)
   • Use case: Second-line after initial screening

3.5. Comparison of the algorithms of support vector machine, convolutional neural network, VGG16, ResNet-50, InceptionV3 and Xception in the analysis of their advantages and disadvantages

SVM only employs support vectors to make predictions, saving memory and computing resources, particularly on large datasets. Considering the intensive optimization process, SVM performance is highly dependent on the aforementioned process. Convolutional Neural Network (CNN) is capable of recognizing patterns, such as shapes and edges, regardless of their position within the image. Meanwhile, VGG16 has a simple and comprehensible structure, with consistent utilization of 3x3 convolutional layers. Despite its simplicity, VGG16 can achieve high performance in image classification tasks. Furthermore, ResNet-50 employs residual connections enabling extensive network training without degradation problems. It can also achieve high performance in various computer vision tasks, while ResNet-50 has the disadvantage in which its deep and complex architecture can require significant computational resources. InceptionV3 is designed for computational efficiency, thereby enabling accelerated training and reduced resource usage. However, the complex design of InceptionV3 can make implementation and maintenance more challenging. To achieve an optimal performance, proper hyperparameter selection and tuning are required. Xception replaces the Inception module with separable convolution, improving efficiency and performance, thus achieving better performance in comparison to InceptionV3 on large datasets. Nevertheless, Xception requires large datasets for effective training. A more complex architecture can increase computational and memory requirements [29]. Tables 2 and 3 presents a comparison of Support Vector Machine, the performance of the Convolutional Neural Network (CNN), VGG16, ResNet-50, InceptionV3, and Xception algorithms.

3.6. Potential limitations or challenges in detecting heart disorders

Data Quality and Availability The problem with many medical data is that they are incomplete, inconsistent, or noisy, which can result in detection models (including AI-based ones) producing inaccurate predictions. ECG or echocardiography signals are difficult to interpret due to noise or artifacts, which can increase the risk of false positives/negatives. The process of labeling medical data requires expertise and time, so datasets are often limited, which can result in models not being sensitive enough to detect rare cases.

4. Conclusion

This study presents a comparative analysis of various deep learning and machine learning models for the classification of cardiac disorders using electrocardiogram (ECG) data. Experimental results show that all models achieved relatively strong performance, with an average accuracy of 78.62% ± 2.81 and an AUC-ROC of 0.8286 ± 0.0209, indicating strong discriminatory ability. Among the evaluated models, Inception-V3 achieved the best overall performance with an accuracy of 82.61%, a recall of 0.6364, and an F1 score of 0.7241, demonstrating a strong balance between sensitivity and precision. Although SVM produced the highest precision (0.933) and specificity (0.9831), its recall performance (0.4242) was limited, indicating reduced sensitivity to positive cases. This study investigates the finding that deep learning architectures, specifically Inception-V3 and Xception, outperform traditional machine learning methods such as SVM in capturing complex ECG patterns associated with cardiac disorders. This demonstrates the potential of deep learning approaches for more accurate and reliable automated heart disease classification. Future research can focus on optimizing model architecture, implementing hybrid ensemble strategies, and integrating larger and more diverse datasets to further improve generalizability and clinical applicability.

Acknowledgement

We would like to thank Serang Raya University for their support, especially the Head of LPPM. We also extend our gratitude to the Insan Unggul College of Computer Science and Technology (STTIKOM), Cilegon, Indonesia, for their support, enabling this research to proceed smoothly.

References

1. Sumiati, et al., *Certainty Cognotive Map (CCM) for Assessing Cognitive map causality using certainty factors for caidiac failure,* ICIC Express, 15 (2021).
2. Sumiati, et al., *Classification of cardiac disorders based on electrocardiogram data with fuzzy cognitive map (FCM) Algoritm approach*, 15 (2021).
3. Sumiati, et al., *Expert system for heart disease based on electrocardiogram data using certainty factor with multiple rule,* IAES International Journal of Artificial Intellegence (IJAI) 10 (2021).
4. R. A. Rajagede et al. *Al-Quran recitation verification for memorization test using Siamese LSTM network*, Commun. Sci. Technol. 6 (2021) 35–40.
5. T. G. Pratama, et al., *Machine learning algorithm for improving performance on 3 AQ-screening classification,* Commun. Sci. Technol. 4 (2019) 44–49
6. M. Abdullah, et al., *Artificial intelligence-based framework for early detection of heart disease using enhanced multilayer perceptron*, 2025.
7. Z. Wang, et al., *Hierarchical deep learning with Generative Adversarial Network for automatic cardiac diagnosis from ECG signals*, Comput. Biol. Med. 155 (2023).
8. P. Wagner, et al., *Explaining deep learning for ECG analysis: Building blocks for auditing and knowledge discovery*, Comput. Biol. Med. 176 (2024) 108525.
9. K. Natarajan, et al., *Efficient Heart Disease Classification Through Stacked Ensemble with Optimized Firefly Feature Selection*, Int. J. Comput. Intell. Syst., 17 (2024) 174.
10. S. Ahmadian, S. M. J. Jalali, S. Raziani, and A. Chalechale, *An efficient cardiovascular disease detection model based on multilayer perceptron and moth-flame optimization*. Expert Syst. 39 (2022) e12914.
11. L. Ali, A. Niamat, J. A. Khan, A. Golliarz, X. Xingzhong, A. Noor, et al., *An optimized stacked support vector machines based expert system for the effective prediction of heart failure,* IEEE Access, 7 (2019) 54007–54014.
12. M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. Quinn, and M. A. Moni, *Heart disease prediction using supervised machine learning algorithms*, performance analysis and comparison. Comput. Biol. Med. 136 (2021) 104672.
13. A. A. Alnuaim, M. Zakariah, P. K. Shukla, A. Alhadlaq, W. A. Hatamleh, H. Tarazi, et al., *Human-computer interaction for recognizing speech emotions using multilayer perceptron classifier*. J. Healthc. Eng. (2022) 6005446.

14. D. Dharmendra, *Prediction of heart failure using support vector machine compared with decision tree algorithm for better accuracy*, in 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS) (Erode: IEEE) (2022).
15. R. Jahangir et al, *ECG-based heart arrhythmia classification using feature engineering and a hybrid stacked machine learning*, BMC Cardiovascular Disorders. 25 (2025) 260.
16. R. Jahangir, S. Ahmed, & M. Malik, *ECG-based cardiac arrhythmias detection using convolutional neural network architectures.* BMC Cardiovasc. Disord., 25 (2025) 150.
17. U. Gupta, P. Sharma, & R. Kaur, *A comprehensive review on efficient artificial intelligence algorithms for ECG analysis,* Expert Systems with Applications, 245 (2024) 123456.
18. A. R. Ismail, S. Q. Nisa, S. A. Shaharuddin, S. I. Masni, & S. A. Suharudin Amin, *Utilising VGG-16 of Convolutional Neural Network for Medical Image Classification*, *International Journal on Perceptive and Cognitive Computing*, 10 (2024) 113–118.
19. R. Hapsari, & A. Purwinarko, *Implementation of Convolutional Neural Network Algorithm Using Vgg-16 Architecture for Image Classification in Facial Images*, Recurs. J. Inform., 1 (2023) 83-92
20. A. Daza, et al., *Stacking ensemble based hyperparameters to diagnosing cardiovascular diseases. Biomedical Signal Processing and Control*. ScienceDirect. (2024)
21. K. Tang, et al., *Optimizing machine learning for enhanced automated ECG interpretation and diagnosis*, Egypt. Inform. J. 28 (2024) 100578.
22. S. Din, et al., *ECG-based cardiac arrhythmia classification through rich feature fusion and hybrid deep learning*,Biomedical Signal Processing and Control. ScienceDirect. (2024)
23. H. Narotamo, M. Dias, R. Santos, A. V. Carreiro, H. Gamboa dan M. Silveira, *Deep learning for ECG classification: A comparative study of 1D and 2D representations and multimodal fusion approaches*, *Biomed. Signal Process. Control* 93 (2024) 106141.
24. E. Westphal, & H. Seitz, *A Machine Learning Method for Defect Detection and Visualization in Selective Laser Sintering based on Convolutional Neural Networks*, Computers, Materials & Continua, 68 (2021) 1915–1933.
25. S. U. Haq, et al., *Reseek-Arrhythmia: Empirical evaluation of ResNet architectures for ECG arrhythmia detection. Sensors,* MDPI 23 (2023) 3891.
26. H. Tsuji & I. Shiojima, *Increased Incidence of ECG Abnormalities in the General Population During the COVID-19 Pandemic.* Int. Heart J., 63 (2022) 678-682.
27. C. F. Gonçalves dos Santos & J. P. Papa, *Avoiding Overfitting: A Survey on Regularization Methods for Convolutional Neural Networks,* (2022).
28. V. Vardana, et al., *A Comprehensiive Review on Heart Disease Risk Prediction using Machine learning and Deep Learning Algorithms*, Arch. Comput. Methods Eng., 32 (2024) 1763-1795.
29. Md Nahid Hasan et al. (2025), *An ensemble based lightweight deep learning model for the prediction of cardiovascular diseases from electrocardiogram images*, Eng. Appl. Artif. Intell., 141 (2025) 109782.