Alsayat *et al. Journal of Big Data* (2025) 12:7
https://doi.org/10.1186/s40537-025-01070-4

Journal of Big Data

RESEARCH

Open Access

Enhancing cardiac diagnostics: a deep learning ensemble approach for precise ECG image classification

Ahmed Alsayat[1*], Alshimaa Abdelraof Mahmoud[2], Saad Alanazi[1], Ayman Mohamed Mostafa[3*], Nasser Alshammari[4], Majed Abdullah Alrowaily[1], Hosameldeen Shabana[5,6] and Mohamed Ezz[1]

*Correspondence:
Ahmed Alsayat
asayat@ju.edu.sa
Ayman Mohamed Mostafa
amhassane@ju.edu.sa
[1]Department of Computer Science, College of Computer and Information Sciences, Jouf University, Sakaka 72388, Saudi Arabia
[2]Department of Information Systems, MCI Academy, Cairo, Egypt
[3]Information Systems Department, College of Computer and Information Sciences, Jouf University, Sakaka 72388, Saudi Arabia
[4]Department of Computer Engineering & Networks, College of Computer and Information Sciences, Jouf University, Sakaka 72388, Saudi Arabia
[5]College of Medicine, Shaqra University, Shaqra, KSA, Saudi Arabia
[6]Faculty of Medicine, Al Azhar University, Cairo, Egypt

Abstract
Cardiovascular diseases are a global health challenge that necessitates improvements in diagnostic accuracy and efficiency. This study examines the potential of deep learning (DL) models for the classification of electrocardiogram (ECG) images to assist in the identification of various cardiac conditions. We initiated a two-tiered experimental framework to investigate the effectiveness of several neural network architectures in this medical application. In the first experiment, eight distinct neural network models were selected based on their top-5 accuracy on the ImageNet validation dataset and were fine-tuned using transfer learning techniques. These models were assessed using a cross-validation scheme, focusing on balanced accuracy, precision, recall, and the F1-score to evaluate their classification capabilities across four cardiac conditions: Myocardial Infarction (MI), abnormal heartbeat, historical MI, and normal ECG patterns. The second experiment extended our inquiry into the power of ensemble learning. By testing all possible combinations of the chosen models, we explored 120 ensemble configurations. The resulting analysis identified the best-performing ensemble set, which did not include the least effective model based on F1 score rankings. The most effective ensemble, composed of Inception, MobileNet, and NASNetLarge, achieved an F1 score of 0.9651 and a balanced accuracy of 0.9640, indicating a superior predictive performance. The ROC curve analysis yielded near-perfect Area Under the Curve (AUC) values for all classes, underscoring the ensemble's proficiency in distinguishing between the specified cardiac conditions. The outcomes of this research highlight the synergistic benefit of ensembles in DL applications for medical imaging and suggest a promising approach for the early detection and diagnosis of cardiac diseases, potentially improving clinical outcomes and patient care.

Keywords Cardiovascular diseases, Deep learning, ECG classification, Neural network architectures, Transfer learning, Ensemble learning

Introduction

Cardiovascular diseases (CVDs) are the leading cause of death globally, presenting a significant challenge for early diagnosis and treatment. The ECG is a pivotal diagnostic tool for the assessment of cardiac function and the identification of pathological conditions [1]. However, manual interpretation of ECGs is subject to variability and depends heavily on the expertise of the practitioner. The ECG has had a massive impact on medicine and the quality of life of many patients. Therefore, the role of AI in ECG is inciting a change [2]. With the advent of machine learning and particularly DL, there is a significant potential to augment diagnostic accuracy and alleviate the workload of medical professionals. DL plays a revolutionizing role to enhance ECG analysis for cardiovascular health. It broadens the scope for EGC to be applied through the development of advanced models that are able to predict disease outcomes, which current prove to be very popular [3]. Recent advances in convolutional neural networks (CNNs) have demonstrated remarkable success in image recognition tasks, suggesting a promising avenue for the application of these models in the interpretation of ECG images. The vast and diverse architectures available today offer an arsenal of tools for tackling complex pattern recognition problems. This study investigates the applicability and performance of various CNN architectures on the task of ECG image classification to assist in the diagnosis of cardiac conditions. Utilizing transfer learning, we adapt pre-trained networks to this domain, hypothesizing that such models can capture the subtle nuances present in ECG tracings. Moreover, the concept of ensemble learning, where multiple models combine their predictive power, provides a compelling strategy to enhance classification performance. This study delves into the construction of various ensemble configurations, seeking to determine the most effective combination of models for the classification of ECG images. We postulate that an ensemble of models will outperform individual architectures due to the complementary learning and generalization capabilities. Our research is driven by the following questions:

1. How effectively can pre-trained neural network models, through transfer learning, be adapted to classify ECG images for the detection of cardiac conditions?
2. What is the comparative performance of different CNN architectures in the classification of ECG images?
3. Can ensemble-learning methods enhance the accuracy of ECG image classification, and what is the optimal ensemble configuration for this task?

The answers to these questions hold profound implications for the field of medical image analysis and for the future of cardiac care. This study not only aims to contribute to the advancement of automated diagnostic methods but also to explore the bounds of artificial intelligence in supporting and enhancing clinical decision-making processes. Through rigorous testing and validation of DL models on a well-curated ECG dataset, we aim to bridge the gap between technological potential and clinical utility.

The integration of AI in healthcare has opened new frontiers in diagnostics, with DL poised at the forefront due to its ability to discern patterns indiscernible to the human eye. In the realm of cardiology, the interpretation of ECGs remains a cornerstone in the diagnosis of CVDs. Accurate and timely analysis of ECGs is crucial, as delays or misinterpretations can lead to adverse patient outcomes. The application of DL models to automate and refine ECG interpretation is a burgeoning field of research, promising to enhance diagnostic processes and patient prognosis. Despite the high stakes associated with CVDs, the utilization of automated systems in clinical practice has been limited. This is partly due to the variable performance of these systems in real-world settings, a gap that this research aims to address. By systematically evaluating the capabilities of different CNN architectures, we can identify models that are not only theoretically sophisticated but also practically viable in clinical environments. This study also considers the scalability and efficiency of model deployment in healthcare settings, where computational resources may be limited. By comparing single-model architectures to ensemble approaches, we assess the trade-offs between computational demands and performance gains. In doing so, we provide a nuanced analysis that factors in the practical aspects of implementing DL solutions in cardiology. As we embark on this investigative journey, we place a particular emphasis on the interpretability and explainability of model decisions. This is vital for gaining the trust of medical practitioners and for the eventual integration of these models into clinical workflows. Our research contributes to a growing body of knowledge that not only aims to advance the technical frontiers of machine learning in healthcare but also considers the ethical, practical, and human-centric aspects of its application. In summary, this introduction sets the stage for a comprehensive exploration of DL in the classification of ECG images, with the ultimate goal of advancing cardiac diagnostics through innovative artificial intelligence solutions.

Related work

In the last few years, the combination of AI and medical diagnosis, especially cardiovascular disease detection with the help of the ECG, has shown tremendous breakthroughs. Traditional methods based on manual feature engineering were limited due to signal complexity; but the advent of DL using convolutional neural networks (CNNs) of unparalleled success makes the ECG analysis easier than ever as it directly learns from raw data, surpasses traditional approaches, and reaches the expert level of accuracy. Transfer learning enhances specific medical imaging applications on the pre-trained models, while the ensemble models combine the predictions of several DL models leading to an extraordinary level of diagnostic precision. This transition from manual to artificial intelligence technologies denotes a new scale of accuracy, and this country of efficiency in cardiology with cardiovascular diseases being diagnosed and treated is improved significantly.

As presented in [4], the authors introduced a novel methodology utilizing the Maximal Overlap Wavelet Packet Transform (MOWPT) combined with Fast Compression Residual Convolutional Neural Networks (FCResNet), achieving an accuracy of 98.79%. This approach underscores the importance of integrating time-scale decomposition techniques with DL for ECG arrhythmia classification. An accuracy of 98.79% was achieved, indicating the effectiveness of combining MOWPT with FCResNet for arrhythmia classification. As proposed in [5], the study produced an innovative neural network solution for automatically finding AF, which consists of 11-layer network that is mainly comprised of a combination of convolutional neural network (CNN) and a modified Elman neural network (MENN). The method provides an opportunity to classify the signal starting from the very beginning. The performance of the proposed model was assessed through ten-fold cross-validation utilizing the MIT-BIH AF database. The findings demonstrated outstanding classification performance, with the model achieving an accuracy rate of 97.4%, a sensitivity of 97.9%, and a specificity of 97.1%.

As presented in [6], the authors applied a transfer DL approach using Continuous Wavelet Transform (CWT) to convert 1-D ECG signals to 2-D scalogram images, classified with a deep convolutional neural network (CNN) using AlexNet architecture. The results attained an impressive accuracy of 95.67%, demonstrating the effectiveness of transfer learning and CNNs for cardiac arrhythmia classification. The authors of [7] Utilized Short-Time Fourier Transform (STFT) to convert ECG signals into time-frequency spectrograms, which were then classified using a 2D Convolutional Neural Network (CNN). They highlighted the utility of transforming ECG signals into time-frequency spectrograms for classification. The study achieved an averaged accuracy of 99.00%, displaying the effectiveness of 2D CNN models in interpreting complex ECG data. The authors of [8] developed DeepECG, which is a system that diagnoses arrhythmias from ECGs using DCNN and transfer learning. The Inception-V3 model within this framework achieved a mean balanced accuracy of 98.46%, highlighting the system's capacity for high-performance ECG arrhythmia classification.

As proposed in [9], the authors conducted a comprehensive review of DL applications in ECG arrhythmia classification, addressing challenges and opportunities for future research. This review emphasized the dominance of CNN models and the potential for advanced denoising and data augmentation techniques to enhance classification accuracy. The review identifies trends, challenges, and opportunities, noting a predominance of CNN models and the need for advanced denoising, data augmentation, and novel DL models for improved classification in clinical settings. The authors of [10] proposed an Image Fusion Model (IFM) that converts heartbeats into images using various techniques before fusion. This approach, leveraging AlexNet CNN, achieved remarkable results, indicating the efficacy of multimodal image fusion in ECG classification. The authors achieved state-of-the-art results in prediction accuracy, precision, and recall, demonstrating the efficacy of multimodal image fusion in ECG classification. The authors of [11] utilize DL technology for the development of a more effective and precise ECG signal classification model. They realized the system on a specially designed CNN (Convolutional Neural Network) called a 1-D convolutional residual neural network (ResNet). This ResNet was constructed to focus on extracting information from the signal using only ECG features and handle the bias in the training data.

As shown in [12], the authors applied MobileNet Convolutional Neural Network architecture on ECG exam images to distinguish between Atrial Fibrillation and Atrial Flutter, incorporating explainable AI methods for model interpretation. They achieved high classification performance, with accuracy of 95.6% and specificity of 99.6%, demonstrating the potential for deep learning in clinical diagnostics. The authors of [13] utilized CNNs for the efficient classification of ECG signals, indicating the effectiveness of deep learning models over conventional machine learning algorithms. The study highlights the potential of deep learning in enhancing the accuracy and efficiency of diagnosing cardiac diseases using ECG data, achieving significant improvements in model performance through parameter optimization. Furthermore, the authors of [14] offers a comprehensive evaluation of various deep learning algorithms applied to ECG diagnosis. The review emphasizes the hierarchical architecture of deep learning, which enables the extraction of higher-level features, leading to enhanced ECG classification accuracy. The review also discusses the strengths and limitations of different deep learning algorithms in ECG diagnosis, suggesting avenues for future research.

As presented in [15], the authors examines the effectiveness of various deep learning models for ECG arrhythmia detection. This study provides insights into the use of deep learning for automatic recognition of arrhythmias, demonstrating deep learning's capacity to surpass the accuracy of manual analysis by experts. The research highlights the significance of model selection and data preparation in achieving high classification performance. In addition, the authors of [16] conducts a thorough review of DL methods applied to ECG signal classification, revealing CNN as the most commonly employed technique. The review underscores the high accuracy achieved by DL models in classifying various arrhythmias, suggesting that DL holds the potential to enhance ECG arrhythmia classification efforts. The authors of [17] explores two multimodal fusion frameworks for enhancing ECG heartbeat classification. The study demonstrates the advantages of combining information from multiple imaging modalities, achieving classification accuracies of over 99% for arrhythmias and myocardial infarction. This innovative approach to ECG classification highlights the potential of multimodal fusion in improving diagnostic accuracy. As presented in [18], the authors investigates the use of 3-D ECG images for cardiac abnormality identification, leveraging the spatial-temporal information in ECG signals. The study presents a novel approach to ECG classification that considers the regional constraints of leads, offering promising classification accuracy in the PhysioNet Challenge 2021. This research underscores the potential of 3-D imaging and CNNs in advancing cardiac diagnostics.

Proposed framework and mechanism

Deep learning (DL) is revolutionizing ECG analysis for CVD diagnosis. Studies have shown that directly using the raw ECG signal outperforms converting it into an image or combining it with other data sources for DL models [19–22]. CLINet, a new DL network achieving high accuracy with just raw ECG data, highlighting DL's potential and paving the way for wearable device integration [23], further supports this. Beyond DL, researchers are exploring other innovative techniques for ECG analysis. One approach combines IoT with deep learning for ECG image classification on a healthcare platform [24]. Another tackles arrhythmia classification by combining a metaheuristic optimization algorithm with machine learning models [25]. A third approach addresses small ECG dataset limitations by employing distant transfer learning, achieving significant accuracy improvements [26]. These advancements collectively demonstrate the potential of various techniques to improve CVD diagnosis and management, leading to more efficient and accessible healthcare solutions.

In this section, we outlined the implementation and outcomes of our proposed methodology for classifying ECG images across various cardiac conditions. As presented in Fig. 1, by utilizing TensorFlow and Keras, we fine-tuned eight pre-trained models on our dataset, employing data augmentation techniques to enhance generalization. Through Experiment 1, we assessed individual model performance, observing varying efficacy in capturing ECG patterns. Experiment 2 involved a model ensemble strategy, where the top-performing models were combined iteratively, resulting in enhanced classification accuracy across all conditions. These findings underscore the efficacy of transfer learning

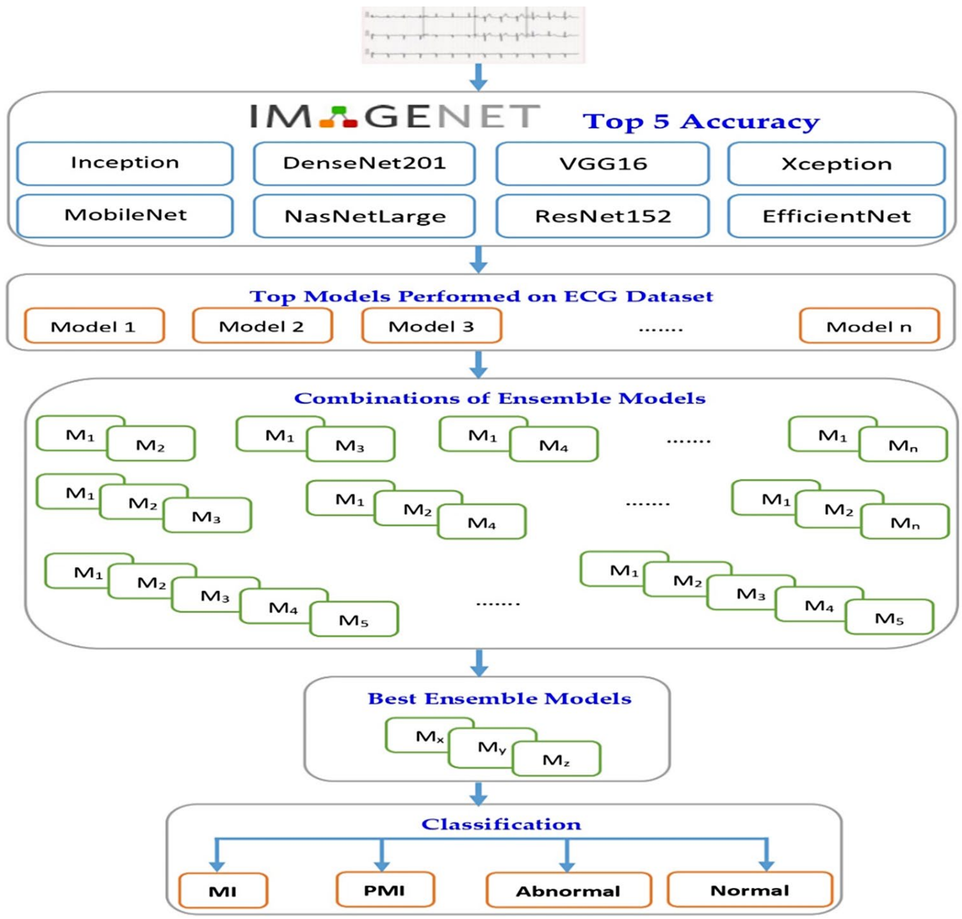

Fig. 1 Proposed ECG ensemble framework

and ensemble techniques in improving cardiovascular disease diagnosis via ECG analysis, with implications for advancing clinical practice and patient care.

Dataset description

The dataset utilized in our study comprises ECG images specifically curated for research on cardiac patients with the primary aim of supporting and advancing scientific research in the field of cardiovascular diseases [27]. The dataset is a significant contribution to the medical and research community, offering a focused resource for developing and testing advanced diagnostic algorithms. The ECG images dataset encompasses a broad spectrum of cardiac conditions, meticulously classified into four primary categories:

1. **Myocardial Infarction Patients**: This group consists of ECG images from patients diagnosed with myocardial infarction (MI), a critical condition caused by the interruption of blood supply to a part of the heart, resulting in heart muscle damage.
2. **Patients with Abnormal Heartbeat**: This category includes ECG recordings from individuals exhibiting arrhythmias or abnormal heart rhythms, which can range from benign to life-threatening conditions.
3. **Patients with a History of MI**: This segment contains ECG images from patients who have previously experienced a myocardial infarction. The inclusion of this group

allows for the examination of long-term cardiac effects post-MI and the identification of recurrent risk patterns.

4. **Normal Persons**: Serving as the control group, these ECG images are from individuals with no diagnosed cardiac abnormalities, providing a baseline for normal heart function and rhythm against which pathological conditions can be contrasted.

The research aims to use a dataset of ECG images categorized into four heart conditions (normal, abnormal heartbeat, myocardial infarction, and previous MI) to develop a classification system. This system, built using transfer learning, would not only differentiate healthy hearts from unhealthy ones but also identify specific heart abnormalities. Early detection and classification of these conditions are crucial for better patient outcomes and reduced healthcare costs. Table 1 explores the overall description of ECG dataset.

Transfer learning models

In this section of our study, which focuses on applying eight transfer learning for the recognition of electrocardiogram (ECG) patterns in cardiac patients, we meticulously selected eight distinct models to ensure a comprehensive evaluation across various neural network architectures. This selection enhances diagnostic accuracy and efficiency. The transfer learning models are presented as follows:

- **VGG16** - Selected for its renowned simplicity and depth [28], offering a foundational baseline within convolutional neural network architectures. It is pivotal for understanding the performance of relatively straightforward, deep architectures in ECG pattern recognition.
- **DenseNet201** - Chosen for its innovative approach to connectivity patterns within the network [29], facilitating improved flow of information and gradients throughout the network, which is crucial for capturing subtle patterns in ECG data.
- **Xception** - Integrated for its advanced use of depthwise separable convolutions, offering a balance between parameter efficiency and computational performance [30]. This model is particularly relevant for distinguishing complex ECG signals with minimal computational resources.
- **MobileNet** - Included for its lightweight architecture for mobile environments, making it an excellent candidate for real-time ECG analysis in resource-constrained settings [31].
- **ResNet152** - Selected for its deep residual learning framework, enabling the training of exceptionally deep networks [32]. Its capability to mitigate the vanishing gradient problem is vital for learning detailed features in ECG data.
- **InceptionResNetV2** - This model combines the Inception architecture's efficiency in handling varied signal scales with ResNet residual connections, offering a potent solution for capturing intricate ECG signal variations [33].

Table 1 ECG dataset description

Class	Count
Normal	284
Myocardial Infarction (MI)	239
History of Myocardial Infarction (HMI)	172
Abnormal Heartbeat	233
Total	**928**

- **NASNetLarge** - Chosen for its state-of-the-art performance, derived from neural architecture search techniques, providing an advanced model capable of capturing and analyzing complex ECG patterns with high accuracy [34].
- **EfficientNetB7** - Integrated for its scalability and balance across different dimensions of the network (width, depth, and resolution), which is instrumental in processing ECG signals with varying degrees of complexity and detail [35].

By leveraging the distinct advantages and innovative features of these eight models, our study aims to explore the potential of transfer learning in enhancing the accuracy and efficiency of ECG pattern recognition for cardiac patients. This diverse set of models not only allows for a comprehensive evaluation of current deep learning approaches in medical diagnostics but also ensures that the findings of our research are robust, scalable, and applicable across a range of clinical settings.

Proposed methodology

Our research is dedicated to the detection of cardiovascular diseases using ECG images, a critical endeavor aimed at leveraging advanced deep learning techniques to improve diagnostic accuracy and efficiency. The methodology is structured around a series of experiments designed to refine and enhance the performance of deep learning models in classifying various cardiac conditions. These conditions include myocardial infarction, abnormal heartbeat patterns, a history of myocardial infarction, and normal heart function.

Experiment 1: model evaluation through transfer learning

The initial phase of our experimental setup involves the application of transfer learning techniques [36] across a selection of eight pre-trained models. The chosen models represent a diverse array of neural network architectures, carefully selected for their potential in image-based classification tasks. For each model, the classification layer is replaced with a new layer tailored to differentiate among the four specified cardiac conditions. The performance of each adapted model is evaluated using cross-validation techniques. This approach allows us to assess the models' generalizability and robustness across different subsets of the dataset, ensuring that our findings are not biased by particular data partitions.

Experiment 2: model ensemble strategy

Following the individual evaluation, we proceed to an ensemble phase where the models demonstrating the highest performance are combined. This ensemble strategy [37, 38] involves pairing the two best-performing models from the initial experiment and evaluating their collective performance. The process is iteratively repeated, incorporating additional models into the ensemble until a decline in performance is observed. Through this method, we aim to identify the optimal ensemble configuration that maximizes diagnostic accuracy.

Training methodology and evaluation Metrics

Our research employs a sophisticated training methodology tailored to enhance the detection of cardiovascular diseases using ECG images. This involves a strategic application of transfer learning on pre-selected models, followed by an ensemble strategy to synergize the strengths of the best-performing models. This section updates the training approach and introduces a comprehensive set of evaluation metrics designed to provide a holistic assessment of model performance.

Training approach

The training process adheres to a structured protocol, initiating with the adaptation of pre-trained models through transfer learning. This entails modifying the models' final classification layers to identify four distinct cardiac conditions accurately. To ensure the models are well-optimized and generalize effectively across diverse data representations, we employ cross-validation coupled with a strategic early stopping mechanism based on model performance. The key training parameters are as follows:

- **Early Stopping Criterion**: Focused on enhancing model generalizability and preventing overfitting. The training process is halted if no improvement is observed in the model's performance metrics over a span of 30 epochs.
- **Epochs**: The maximum number of training cycles is capped at 100, balancing between adequate learning time and preventing overtraining.
- **Batch Size**: Set at 16, this parameter optimizes the computational efficiency and stability of the gradient updates during training.

Evaluation metrics

To assess the models' diagnostic capabilities, we employ a multi-faceted evaluation framework that includes balanced accuracy, precision, recall, and multi-class ROC-AUC scores. These metrics are calculated for each experiment to ensure a thorough examination of the models' performance.

- **Balanced Accuracy**: This metric offers a more nuanced view of accuracy, especially important in our dataset where class distributions might be uneven. It calculates the average accuracy obtained on all classes, ensuring that each class contributes equally to the final score.
- Precision and Recall: These metrics provide insights into the models' ability to identify positive cases and the proportion of actual positive cases correctly identified, respectively. Precision is crucial for minimizing false positives, while recall is key to reducing false negatives — both critical in the medical diagnosis context.
- Multi-Class ROC-AUC: The ROC curve and the AUC for multi-class scenarios extend the binary classification analysis to our multi-class problem. This metric provides a comprehensive view of the models' ability to distinguish between all classes across various threshold levels, offering a robust measure of model performance that is unaffected by class imbalance.

By leveraging these evaluation metrics, we ensure a balanced and thorough analysis of each model's and ensembles diagnostic performance. This methodical approach allows us to identify the most effective configurations for ECG image classification, ultimately aiming to advance the detection and understanding of cardiovascular diseases.

Table 2 Confusion matrix for Inception model

Category	MI	Abnormal	HMI	Normal
MI	**1.00**	0.00	0.00	0.00
Abnormal	0.01	**0.93**	0.04	0.02
HMI	0.00	0.05	**0.91**	0.04
Normal	0.00	0.01	0.01	**0.98**

Table 3 Classification report for Inception model

Performance	Precision	Recall	F1-Score
MI	0.99	1.00	0.99
Abnormal	0.95	0.93	0.94
HMI	0.93	0.91	0.92
Normal	0.96	0.98	0.97
Avg/Total	**0.96**	**0.96**	**0.96**

Result and discussion

Experiment 1: evaluation of eight top deep learning models

In Experiment 1, our objective was to assess the performance of eight deep learning models, adapted via transfer learning, for the classification of ECG images into four categories: Myocardial Infarction (MI), abnormal heartbeat patterns, patients with a history of MI (HMI), and normal. The models were evaluated based on precision, recall, f1-score, balanced accuracy, and multi-class ROC-AUC scores. Below, we provide a detailed breakdown of each model's performance, including classification reports, confusion matrices, and a description of ROC curves.

Inception model

The Inception model's performance in classifying ECG images into four categories: MI, AH, HMI, N, is quantitatively assessed through its confusion matrix and classification report. As presented in Table 2, the confusion matrix demonstrates the Inception model's strong ability to classify MI and Normal cases with almost perfect accuracy. It also shows respectable performance in distinguishing Abnormal and HMI cases, albeit with slight confusion among these classes. As presented in Table 3, the classification report further underscores the Inception model's high precision, recall, and F1-score across all classes, with an overall average that highlights its efficacy in ECG image classification. The model excels particularly in identifying MI and Normal conditions, achieving near-perfect scores. Its performance with Abnormal and HMI conditions, while slightly lower, remains robust, demonstrating the model's comprehensive learning from the ECG images. The Inception model displaying a promising application for ECG image classification, particularly in distinguishing between critical cardiac conditions with high accuracy. Its ability to perform well across diverse conditions suggests its potential utility as a supportive tool in cardiac diagnostics, offering a reliable automated method to assist medical professionals in their evaluations.

As shown in Fig. 2, the ROC curve for the Inception model displays three nearly perfect lines for the classes myocardial infarction (MI), abnormal, and normal ECGs, each with an AUC very close to 1.00. This suggests that the model has outstanding classification performance for these conditions, being able to distinguish the presence of these conditions from their absence with high accuracy. The HMI (History of Myocardial Infarction) class also shows excellent performance, with an AUC of 0.97, indicating a

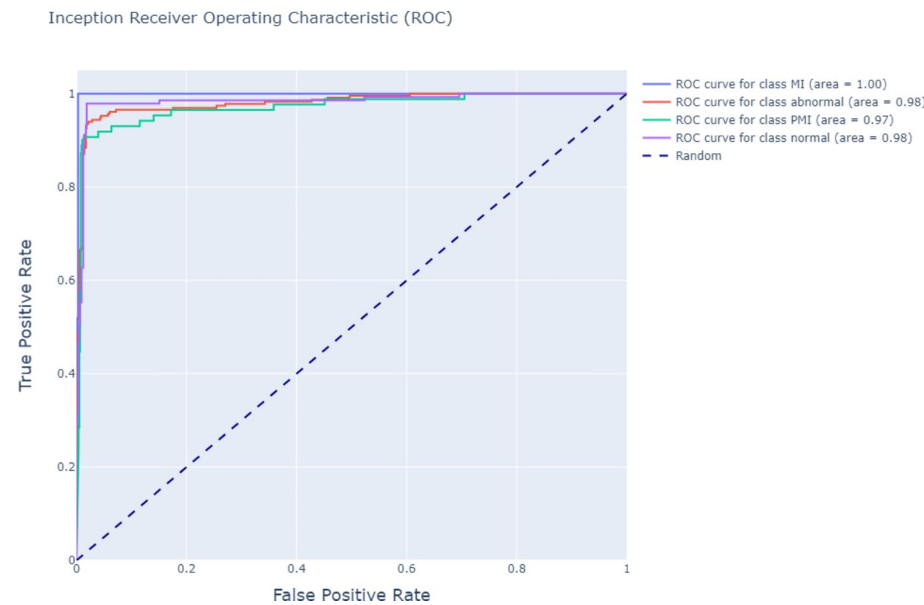

Fig. 2 The ROC curve for inception model

Table 4 Confusion matrix for DenseNet201 model

Category	MI	Abnormal	HMI	Normal
MI	**1.00**	0.00	0.00	0.00
Abnormal	0.01	**0.91**	0.05	0.03
HMI	0.00	0.06	**0.92**	0.02
Normal	0.01	0.02	0.02	**0.95**

high true positive rate and a low false positive rate for this condition. Overall, the ROC curve demonstrates that the Inception model is highly effective at correctly classifying the different ECG classes without being confused by false positives.

DenseNet201 model

The DenseNet201 model's classification performance for ECG image categorization presents promising results, especially in identifying instances of myocardial infarction (MI) and differentiating normal ECG patterns. As presented in Table 4, the confusion matrix reflects the model's strong ability to correctly classify cases of MI with perfect precision and high success rate for Normal class. However, the model appears to have some difficulty distinguishing between cases of abnormal ECG patterns and those with a history of myocardial infarction (HMI), as indicated by the off-diagonal values in the confusion matrix.

The classification report presented in Table 5 indicates high precision and recall across all categories, with the model performing exceptionally well in identifying normal ECG patterns and MI cases. While the performance on abnormal and HMI categories is slightly lower, it remains robust, with high precision and recall indicating reliable diagnostic capability. DenseNet201 exhibits a strong potential for ECG image classification, excelling particularly in the accurate identification of normal cardiac function and myocardial infarction. The model's high precision and recall suggest that it could be a valuable asset in assisting medical professionals with ECG interpretation. Its slight

Table 5 Classification report for DenseNet201 model

Performance	Precision	Recall	F1-Score
MI	0.98	1.00	0.99
Abnormal	0.87	0.91	0.89
HMI	0.95	0.92	0.93
Normal	0.97	0.95	0.96
Avg/Total	**0.94**	**0.95**	**0.94**

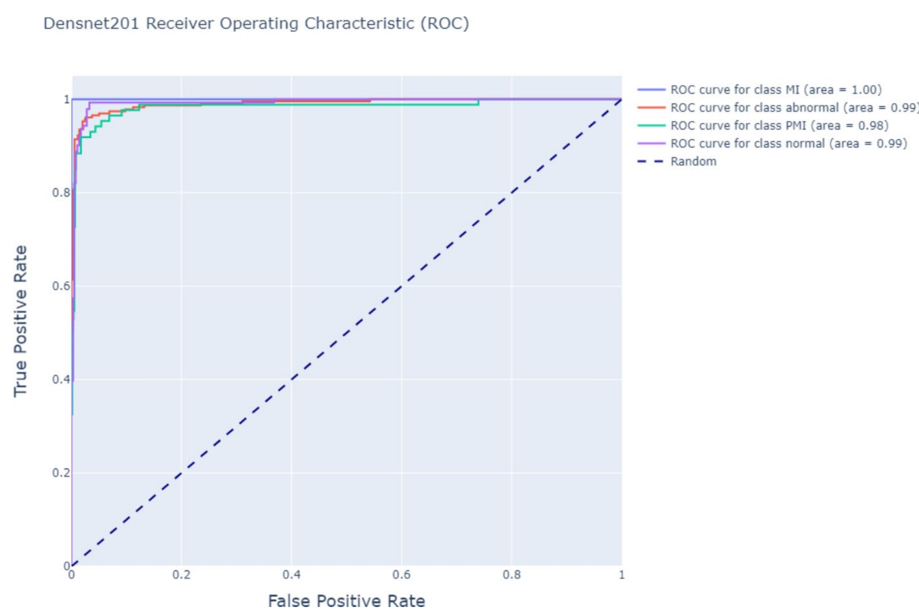

Fig. 3 The ROC curve for DenseNet201 model

challenges with certain classes point to areas for further refinement, potentially through additional training data or ensemble strategies to bolster its diagnostic power.

As presented in Fig. 3, the ROC curve for the DenseNet201 model demonstrates excellent performance across all classes. Each line on the plot represents one of the classes (MI, Abnormal, HMI, and Normal), and the closer the curve is to the top-left corner, the better the model is at distinguishing between that particular class and the rest. The AUC for MI is perfect at 1.00, indicating outstanding model performance with maximum sensitivity and specificity—signifying that it can distinguish MI from other classes without any false positives or negatives. Similarly, the AUCs for Abnormal and Normal classes are both at 0.99, displaying a very high level of discriminative power. The HMI class has a slightly lower AUC at 0.98, which still reflects a high capability of the model to correctly classify HMI cases. Overall, the attached ROC curve illustrates that the DenseNet201 model is highly effective at ECG image classification, performing exceptionally well across all considered cardiac conditions.

VGG16 model

The VGG16 model's proficiency in ECG image classification is reflected through its high precision and recall rates, especially in recognizing cases of myocardial infarction (MI) and normal ECG patterns. This confusion matrix presented in Table 6 indicates a strong correct classification rate for MI and Normal cases with virtually no misclassifications. It suggests some confusion between Abnormal and HMI classes, albeit relatively minor.

Table 6 Confusion matrix for VGG16 model

Category	MI	Abnormal	HMI	Normal
MI	**1.00**	0.00	0.00	0.00
Abnormal	0.01	**0.93**	0.03	0.03
HMI	0.00	0.07	**0.90**	0.03
Normal	0.00	0.02	0.01	**0.97**

Table 7 Classification report for VGG16 model

Performance	Precision	Recall	F1-Score
MI	0.98	1.00	0.99
Abnormal	0.96	0.93	0.95
HMI	0.94	0.90	0.92
Normal	0.94	0.98	0.96
Avg/Total	**0.96**	**0.96**	**0.96**

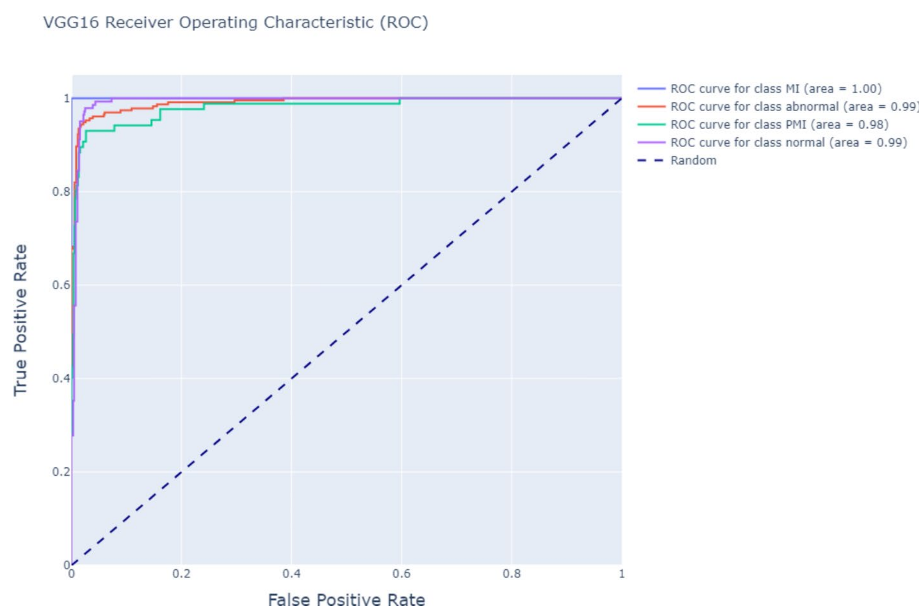

Fig. 4 The ROC curve for VGG16 model

The classification report presented in Table 7 illustrates the model's excellent precision and its ability to recall the true positive cases effectively, yielding a high F1-score. The VGG16 shows a particularly strong performance in identifying MI, which is essential for clinical applications where accurate detection of such conditions is crucial. The VGG16 model has demonstrated itself to be a potent tool for the classification of ECG images, with its capability to achieve high accuracy across all classes. The model's precision and recall scores suggest that it has effectively learned the distinguishing features of various cardiac conditions from the ECG data, making it a valuable asset for aiding in the diagnosis and study of cardiovascular diseases.

The ROC curve for the VGG16 model presented in Fig. 4 indicates that it performs with high accuracy in classifying ECG images. The model exhibits near-perfect classification for myocardial infarction (MI) with an AUC of 1.00, suggesting it can distinguish MI cases from non-MI cases almost flawlessly. The curves for abnormal ECGs and normal ECGs show excellent performance, with both having an AUC of 0.99. The HMI

Table 8 Confusion matrix for Xception model

Category	MI	Abnormal	HMI	Normal
MI	**1.00**	0.00	0.00	0.00
Abnormal	0.02	**0.90**	0.05	0.03
HMI	0.01	0.08	**0.92**	0.01
Normal	0.00	0.01	0.02	**0.97**

Table 9 Classification report for Xception model

Performance	Precision	Recall	F1-Score
MI	0.97	1.00	0.98
Abnormal	0.99	0.90	0.94
HMI	0.92	0.92	0.92
Normal	0.94	0.99	0.96
Avg/Total	**0.96**	**0.96**	**0.96**

(previous myocardial infarction) class has a slightly lower AUC at 0.98, yet this is still indicative of a very strong ability to correctly classify cases. The AUC values close to 1.00 across all classes demonstrate the VGG16 model's robustness in ECG classification tasks, confirming its potential usefulness in clinical settings for supporting the diagnosis of various cardiac conditions.

Xception model

The Xception model was subjected to rigorous testing in our ECG image classification task, yielding results in distinguishing various cardiac conditions. The confusion matrix proposed in Table 8 illustrates an exceptional performance in correctly identifying MI cases, and a very high success rate for Normal class identification. However, there is a slightly increased rate of misclassification between the Abnormal and HMI classes, as compared to the MI and Normal categories. The classification report presented in Table 9 underscores the model's accuracy, with a high level of precision and recall demonstrated across all classes. Notably, the precision for the Abnormal class is remarkable, suggesting that the Xception model is particularly adept at identifying true positive cases in this category. The Xception model has validated its efficacy in ECG image classification, displaying an excellent balance of precision and recall. Its high F1-scores across all cardiac conditions, particularly in the Abnormal class, highlight its potential as a diagnostic tool in cardiology. The robustness of the Xception model, as evidenced by these results, points towards its suitability for aiding clinicians in the rapid and accurate diagnosis of cardiac conditions using ECG data.

The ROC curve for the Xception model in Fig. 5 shows a high-level performance in ECG image classification. Each of the lines represents a different class — MI, abnormal, HMI, and normal — and all are seen hugging the top-left corner of the plot, indicating excellent classification ability. For the class MI, the AUC is perfect at 1.00, demonstrating that the Xception model can identify MI with maximum sensitivity and specificity. Similarly impressive is the model's performance on abnormal ECGs and normal ECGs, with AUC values of 0.99 and 0.99, respectively. The HMI class shows an AUC of 0.99, which, while not perfect, is still indicative of a very strong performance. These AUC scores close to 1 suggest that the Xception model is extremely capable in distinguishing between the different categories of ECG images, providing a reliable tool that could potentially be used in clinical settings for the diagnosis of cardiac conditions.

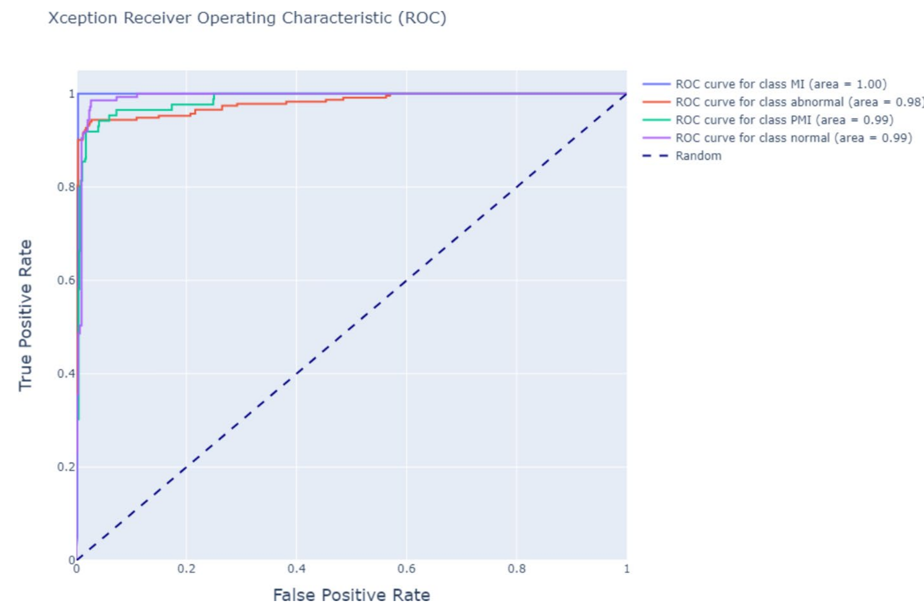

Fig. 5 The ROC curve for Xception model

Table 10 Confusion matrix for MobileNet model

Category	MI	Abnormal	HMI	Normal
MI	**1.00**	0.00	0.00	0.00
Abnormal	0.03	**0.92**	0.03	0.02
HMI	0.02	0.06	**0.93**	0.01
Normal	0.01	0.03	0.02	**0.94**

Table 11 Classification report for MobileNet model

Performance	Precision	Recall	F1-Score
MI	0.97	1.00	0.99
Abnormal	0.98	0.92	0.95
HMI	0.89	0.93	0.91
Normal	0.95	0.95	0.95
Avg/Total	**0.95**	**0.95**	**0.95**

MobileNet model

The MobileNet model was examined for its performance in classifying ECG images, an essential part of automating the diagnosis of cardiac conditions. As presented in Table 10, the confusion matrix indicates a high level of accuracy in classifying MI and Normal cases, with the model showing a good ability to distinguish between the various cardiac conditions. It exhibits some misclassifications within the Abnormal and Normal classes, suggesting areas where the model may benefit from further tuning. As explained in Table 11, the model's precision and recall are commendable across all categories, with an overall F1-score reflecting its effectiveness in ECG classification. MobileNet demonstrates a particularly strong capability to identify cases of MI, a critical requirement for clinical applications. MobileNet proves to be an effective model for ECG image classification, striking a balance between accuracy and computational efficiency—a key consideration for real-time diagnostic systems. While it shows strong performance, particularly in recognizing MI and HMI conditions, the slight confusion

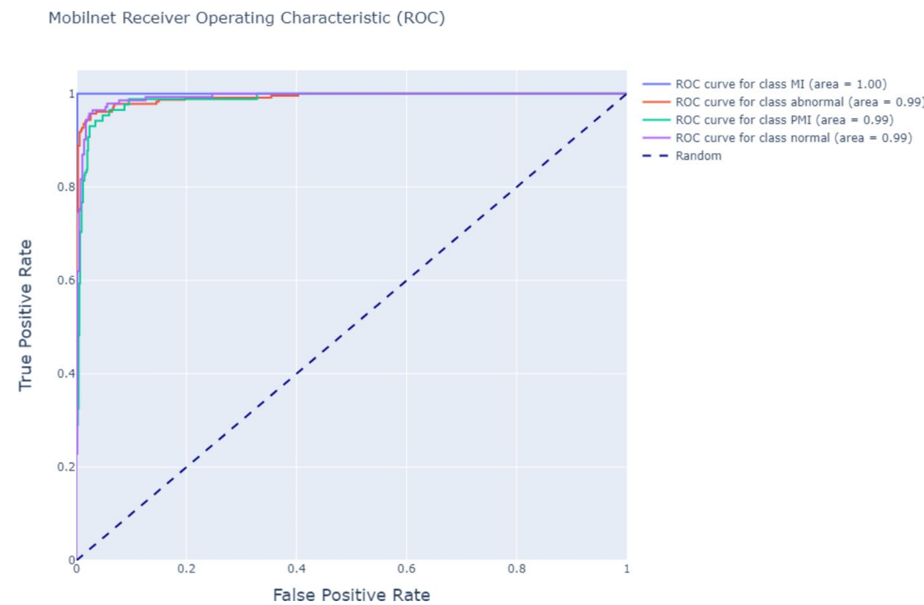

Fig. 6 The ROC curve for MobileNet model

Table 12 Confusion matrix for NASNetLarge model

Category	MI	Abnormal	HMI	Normal
MI	**1.00**	0.00	0.00	0.00
Abnormal	0.00	**0.91**	0.06	0.03
HMI	0.00	0.05	**0.93**	0.02
Normal	0.01	0.04	0.03	**0.92**

in the Normal and Abnormal categories may guide future optimizations. These results highlight MobileNet's potential as a diagnostic aid, providing a robust option for quick and reliable cardiac assessment.

As presented in Fig. 6, the ROC curve for the MobileNet model indicates that it performs with excellence in classifying ECG images. The AUC for MI is 1.00, showing that the model can distinguish MI from non-MI cases with perfect sensitivity and specificity. Similarly, the AUC for the abnormal and normal classes are both at 0.99, signifying an outstanding ability to classify these conditions correctly. The HMI class also has a high AUC of 0.99, suggesting that the model can accurately identify cases with a high true positive rate and a low false positive rate. These results demonstrate that MobileNet is highly effective in ECG image classification and could be a valuable tool for assisting healthcare professionals in diagnosing various cardiac conditions.

NASNetLarge model

NASNetLarge is recognized for its architecture's complexity and capacity to handle large-scale image recognition tasks. The Model was tested for its capability in ECG image classification, a vital tool in diagnosing cardiac conditions. The confusion matrix presented in Table 12 explains an impressive capability of the model to identify MI cases. It also displays a high true positive rate for HMI, although with a slightly higher rate of false positives as compared to other models, particularly within the Normal category.

The classification report metrics presented in Table 13 highlight NASNetLarge precision and recall, reflecting its strengths in distinguishing between normal and abnormal

Table 13 Classification report for NASNetLarge model

Performance	Precision	Recall	F1-Score
MI	0.99	1.00	0.99
Abnormal	0.96	0.91	0.93
HMI	0.91	0.93	0.92
Normal	0.94	0.95	0.95
Avg/Total	**0.95**	**0.95**	**0.95**

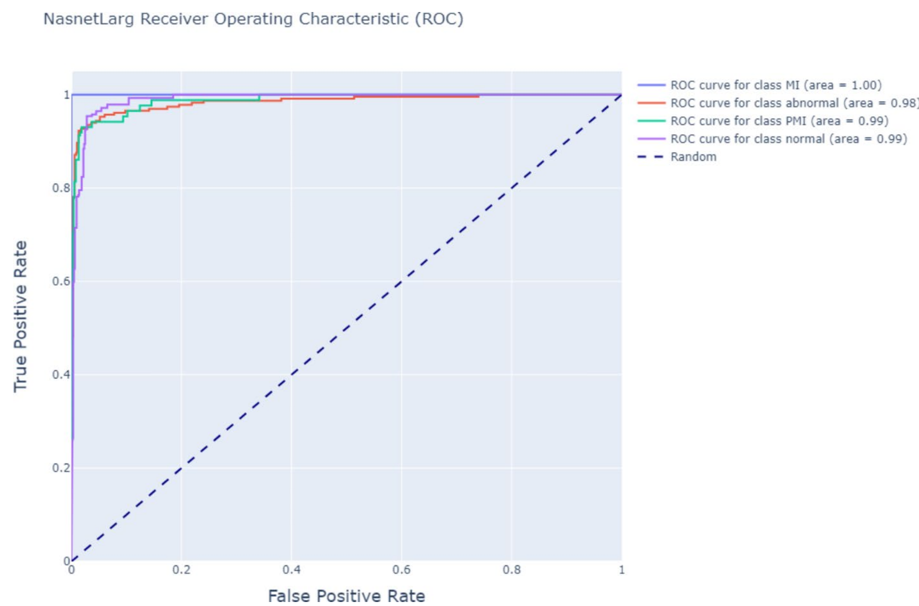

Fig. 7 The ROC curve for NASNetLarge model

cardiac conditions. It achieves a commendable F1-score, emphasizing its balanced performance across the different classes. The NASNetLarge model demonstrates that it is a powerful tool in the realm of medical image analysis, with a particular aptitude for ECG classification. Its high precision and recall across all classes, accompanied by strong F1-scores, indicate that it has the potential to support cardiologists in the diagnostic process. The model's slight tendency to misclassify within the Normal class suggests an area for further refinement, which could involve additional training or the incorporation of NASNetLarge into an ensemble to mitigate such errors. Nonetheless, NASNetLarge stands out as a formidable model for automatic ECG interpretation, providing substantial value to the field of cardiac diagnostics.

As presented in Fig. 7, the ROC curve for the NASNetLarge model shows that it performs exceptionally well in classifying ECG images. The perfect AUC of 1.00 for myocardial infarction (MI) indicates flawless discrimination between MI and non-MI instances. For abnormal ECGs, the AUC is very close to perfection at 0.98, and the model shows a similarly high performance for normal ECGs with an AUC of 0.99. The performance on previous myocardial infarction (HMI) cases is also impressive with an AUC of 0.99. Such high AUC values across all categories confirm that NASNetLarge has a robust capability for ECG image classification, making it a potentially powerful tool in supporting the diagnosis and treatment of various cardiac conditions.

Table 14 Confusion matrix for ResNet152 model

Category	MI	Abnormal	HMI	Normal
MI	**1.00**	0.00	0.00	0.00
Abnormal	0.00	**0.93**	0.04	0.03
HMI	0.00	0.07	**0.87**	0.06
Normal	0.01	0.05	0.07	**0.87**

Table 15 Classification report for ResNet152 model

Performance	Precision	Recall	F1-Score
MI	0.99	1.00	0.99
Abnormal	0.94	0.93	0.94
HMI	0.88	0.87	0.87
Normal	0.92	0.93	0.92
Avg/Total	**0.93**	**0.93**	**0.93**

ResNet152 model

The ResNet152 model, known for its deep residual learning framework, was put to the test for classifying ECG images into distinct cardiac conditions. As proposed in Table 14, the confusion matrix reveals ResNet152's high accuracy in classifying MI cases correctly, while also showing reasonable performance for the abnormal class. However, it demonstrates some limitations with HMI and Normal categories, as indicated by the slightly higher false positive rates. The classification report presented in Table 15 underscores the strengths of ResNet152 in terms of precision and recall, with a particularly strong performance in correctly identifying cases of MI. Nonetheless, the F1-scores for HMI and Normal suggest there is room for improvement, perhaps indicating a need for more nuanced feature extraction or model tuning for these specific classes. ResNet152 has affirmed its capability as a robust model for ECG image classification, offering solid precision and recall values. Despite facing some challenges in distinguishing HMI and Normal cases with the highest accuracy, the model holds considerable promise for supporting cardiology diagnostics. Its performance affirms the potential of deep residual networks in medical imaging, inviting further optimization to enhance its diagnostic accuracy.

Figure 8 explained the ROC curve for the ResNet152 model that indicates its strong performance in ECG image classification. The model achieves an AUC of 1.00 for the MI class, which means it perfectly distinguishes MI cases from non-MI cases without any false positives or negatives. The abnormal and normal classes also exhibit excellent AUC values of 0.99 and 0.98, respectively, demonstrating the model's high capability in correctly identifying these conditions. The performance on HMI cases is also robust with an AUC of 0.97. Overall, these AUC scores show that ResNet152 has high sensitivity and specificity across the different cardiac conditions, marking it as a reliable model for ECG classification in a clinical context.

EfficientNet model

EfficientNet was evaluated for its performance in classifying ECG images into various cardiac conditions. The confusion matrix presented in Table 16 indicates challenges in the model's performance across all classes, with notably lower accuracy for the Normal condition and significant misclassification rates between the classes. This suggests difficulty in distinguishing between the nuanced patterns of cardiac conditions in ECG

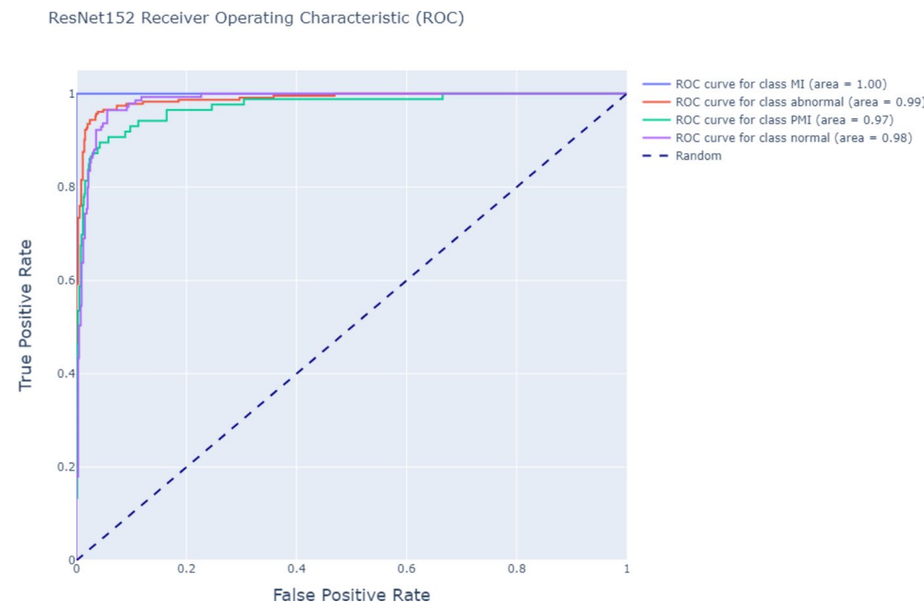

Fig. 8 The ROC curve for ResNet152 model

Table 16 Confusion matrix for EfficientNet model

Category	MI	Abnormal	HMI	Normal
MI	**0.78**	0.07	0.13	0.02
Abnormal	0.14	**0.55**	0.26	0.05
HMI	0.10	0.22	**0.53**	0.15
Normal	0.15	0.12	0.39	**0.34**

Table 17 Classification report for EfficientNet model

Performance	Precision	Recall	F1-Score
MI	0.67	0.78	0.72
Abnormal	0.59	0.55	0.57
HMI	0.31	0.53	0.39
Normal	0.69	0.34	0.46
Avg/Total	**0.56**	**0.55**	**0.53**

images. The classification report in Table 17 reflects lower precision, recall, and F1-score values across all categories when compared to other models tested. This overall performance indicates that EfficientNet, while potent in general image classification tasks, may require further optimization and specific tuning for the complexities of ECG image analysis. The EfficientNet model demonstrated some capacity to classify ECG images but did not achieve the level of accuracy and reliability seen with other models in this study. Its lower performance metrics, particularly in distinguishing Normal conditions and higher misclassification rates, highlight the challenges of applying EfficientNet to ECG image classification without substantial customization. This finding underscores the importance of model selection and the potential need for specialized architectural modifications or training strategies tailored to the unique characteristics of medical imaging data, particularly in cardiology.

As presented in Fig. 9, the AUC for MI class is 0.89, which indicates a good but not perfect ability to distinguish MI from non-MI cases. The curve for the abnormal class

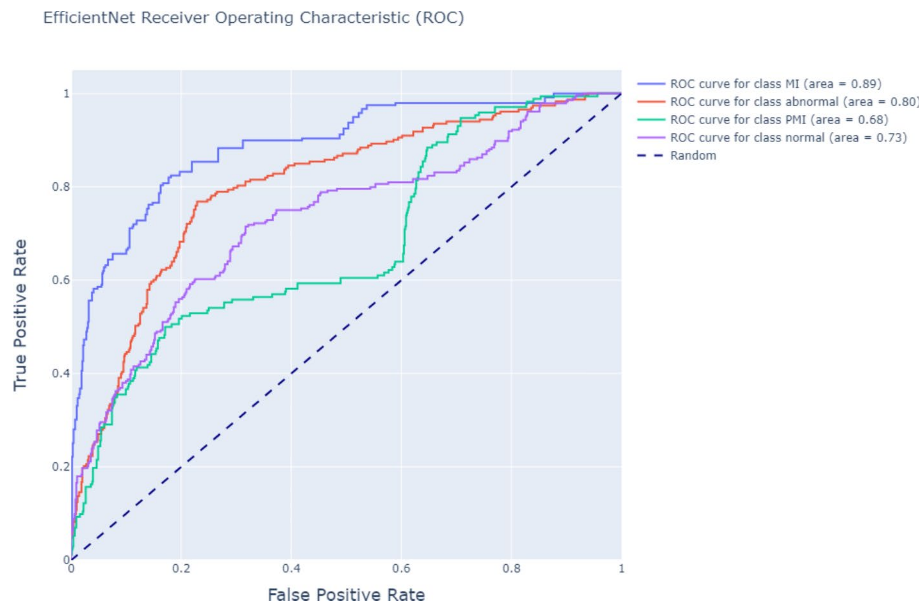

Fig. 9 The ROC curve for EfficientNet model

Table 18 Performance for transfer learning models in experiment 1

Model	Balanced Accuracy	F1 Score	Precision (Avg)	Recall (Avg)
Inception	**95.32%**	**95.43%**	**96%**	**96%**
DenseNet201	95.29%	95.41%	94%	95%
VGG16	95.14%	95.35%	96%	96%
Xception	95.15%	95.27%	96%	96%
MobileNet	94.98%	94.88%	95%	95%
NASNetLarge	94.86%	94.83%	95%	95%
ResNet152	93.07%	93.15%	93%	93%
EfficientNet	54.95%	53.40%	56%	55%

has an AUC of 0.80, suggesting that while the model can identify abnormal cases, there is room for improvement in its sensitivity and specificity. The AUC for the HMI class is lower at 0.68, indicating that the model has difficulty distinguishing these cases accurately. The normal class curve has the lowest AUC at 0.73, reflecting challenges in correctly classifying normal ECG images. These AUC values, which are lower compared to other models discussed earlier, suggest that EfficientNet, while possessing a good predictive capacity, may not perform as strongly as other models in this specific application of ECG image classification. It indicates that EfficientNet might benefit from further optimization or might be more effective when used as part of an ensemble rather than on its own.

To summarize the overall performance for the eight transfer learning models, Table 18 highlights the balanced accuracy and F1 score.

As explained in Table 18, the Precision and Recall columns are provided as averages to give an overall sense of each model's accuracy and sensitivity across the four classes. The Balanced Accuracy and F1 score remain the primary metrics for comparing model performance, reflecting both the accuracy and balance between precision and recall. This summary highlights the significant variance in performance across models, with

Inception, DenseNet201, VGG16, and Xception demonstrating superior capabilities in classifying ECG images accurately.

Based on the summarized data, Inception, DenseNet201, VGG16, Xception, MobileNet, NASNetLarge, and ResNet152 stand out as the top performers, with high balanced accuracy and F1 scores. These models not only excel in individual class predictions but also maintain a strong balance across precision and recall, making them ideal candidates for further analysis and potential ensemble strategies. EfficientNet, despite its general acclaim in various domains, shows a marked underperformance in this particular application, emphasizing the importance of domain-specific model evaluation.

Experiment 2: model ensemble analysis

In Experiment 2, our approach took a more exhaustive route by exploring all possible combinations of ensemble models derived from the initially selected set, excluding only EfficientNet due to its lower performance. This exhaustive analysis resulted in 120 distinct ensemble combinations from which the objective was to ascertain the ensemble set that delivers the highest efficiency. The efficiency will be evaluated based on F1 score for ECG image classification for the categories MI, abnormal, HMI, and normal heart function.

Top 10 ensemble sets

The comprehensive ensemble evaluation led to the identification of the top 10 performing sets, ranked by their F1 scores. As presented in Table 19, the metric was prioritized given its critical balance between precision and recall, which is especially important in medical diagnostic applications where both false positives and false negatives have significant implications.

The selection of ensemble set of deep learning models is based on the following issues:

- **Diversity and Performance**: The top ensemble configurations illustrate a diverse mix of models, with NASNetLarge appearing in every top-performing ensemble. This suggests the model's strong generalization capability and its effectiveness in capturing complex patterns in ECG images.
- **Synergy among Models**: The highest F1 score was achieved by combining Inception, MobileNet, and NASNetLarge, indicating a synergistic effect that leverages the unique strengths of each model to improve classification performance.
- **Impact of Combining Models**: Ensembles incorporating a mix of models, including DenseNet201, VGG16, and Xception, alongside NASNetLarge and MobileNet,

Table 19 Top 10 ensemble set of models

Rank	Ensemble Name	F1 Score
1	Inception, MobileNet, NASNetLarge	0.9651
2	VGG16, MobileNet, NASNetLarge	0.9635
3	Inception, DenseNet201, VGG16, NASNetLarge	0.9631
4	Inception, DenseNet201, Xception, NASNetLarge	0.9628
5	DenseNet201, Xception, NASNetLarge	0.9627
6	VGG16, Xception, MobileNet, NASNetLarge, ResNet152	0.9626
7	Inception, DenseNet201, MobileNet, NASNetLarge	0.9624
8	Inception, DenseNet201, Xception, MobileNet, NASNetLarge	0.9623
9	Inception, VGG16, Xception, MobileNet, NASNetLarge	0.9623
10	Inception, NASNetLarge	0.9622

consistently show high F1 scores, demonstrating the benefit of model diversity in ensemble strategies.

- **Optimal Ensemble Selection**: Despite the slight differences in F1 scores among the top ensembles, the optimal choice would consider computational efficiency and model complexity. The best ensemble, in this case, provides a balance between high classification accuracy and operational feasibility.

The exhaustive ensemble analysis underscores the potential of combining diverse models to enhance the accuracy of ECG image classification. The top-performing ensemble sets, particularly those featuring NASNetLarge, demonstrate the importance of strategic model selection to capitalize on the complementary strengths of different architectures. This approach not only achieves superior classification performance but also offers insights into effective ensemble strategies for complex medical imaging tasks, setting a foundation for future research and application in cardiovascular disease diagnostics.

Bottom 10 ensemble sets

As shown in Table 20, a summary of the ensemble sets that displayed the least performance in terms of F1 score, which indicates a lower balance between precision and recall compared to the other ensemble combinations tested. The following issues explore the main performance of the ensemble models:

- The ensemble with the lowest performance was DenseNet201, MobileNet, and ResNet152, which still achieved a relatively high F1 score of 0.9498, suggesting that even the least effective combinations in this study were still good.
- The presence of ResNet152 in many of the lower-performing ensembles could indicate that this particular model did not synergize as well as others within the ensemble context for this dataset and task.
- The combination of Xception and ResNet152 stands out with the lowest F1 score of 0.9360, which is a notable drop from the others. This particular ensemble might be less capable of capturing the nuanced patterns in ECG images required for accurate classification.

This analysis of the bottom-performing ensembles reveals important insights into the compatibility of different architectures when combined. It highlights that while certain models are strong individually, their performance does not always scale linearly when paired with others in an ensemble. Understanding the dynamics between different

Table 20 Bottom 10 ensemble Set of models

Rank	Ensemble Name	F1 Score
1	DenseNet201, MobileNet, ResNet152	0.9498
2	Inception, Xception	0.9493
3	VGG16, MobileNet	0.9488
4	DenseNet201, Xception	0.9487
5	Inception, ResNet152	0.9483
6	Xception, MobileNet	0.9471
7	VGG16, ResNet152	0.9454
8	NASNetLarge, ResNet152	0.9450
9	DenseNet201, ResNet152	0.9448
10	Xception, ResNet152	0.9360

Table 21 Confusion matrix for best ensemble model

Category	MI	Abnormal	HMI	Normal
MI	**1.00**	0.00	0.00	0.00
Abnormal	0.00	**0.97**	0.03	0.00
HMI	0.00	0.01	**0.92**	0.07
Normal	0.00	0.01	0.01	**0.97**

Table 22 Classification report for best ensemble model

Performance	Precision	Recall	F1-Score
MI	1.00	1.00	1.00
Abnormal	0.97	0.97	0.97
HMI	0.94	0.92	0.93
Normal	0.96	0.97	0.96
Avg/Total	**0.97**	**0.97**	**0.97**

models is crucial for optimizing ensemble strategies, and these findings can direct future research toward more effective combinations.

Best ensemble performance

The ensemble of Inception, MobileNet, and NASNetLarge models marked a significant achievement in our research, setting a new standard for ECG image classification performance. This section details the ensemble's effectiveness, as highlighted by its confusion matrix, classification report, and ROC curve analysis.

Confusion matrix for best ensemble model

The confusion matrix for the best-performing ensemble illustrates nearly flawless classification across all cardiac conditions. As explained in Table 21, the matrix not only underscores the ensemble's capability to correctly identify MI cases with 100% accuracy but also explain its high accuracy in classifying Abnormal, HMI, and Normal conditions with minimal confusion between the classes.

Classification report for best ensemble model

The classification report as depicted in Table 22 reflects a high degree of precision and recall across the board, with impressive F1-scores in each category.

The classification report reaffirms the ensemble's excellence, demonstrating a remarkable balance of precision and recall across all categories. It achieved perfect scores in the MI class, which is critical for detecting potentially life-threatening conditions. The high F1-scores in other categories further validate the ensemble's comprehensive learning capabilities and its robustness in handling diverse cardiac conditions. The ensemble of Inception, MobileNet, and NASNetLarge models emerged as the pinnacle of ECG image classification within this study. By combining the strengths of three powerful architectures, this ensemble has not only achieved exceptional accuracy but also set a benchmark for future research in medical image analysis. Its success illustrates the potential of ensemble learning in enhancing diagnostic precision, offering promising implications for the integration of AI technologies in cardiology. This ensemble strategy paves the way for developing advanced diagnostic tools that can support healthcare professionals in delivering timely and accurate care to patients with cardiovascular diseases.

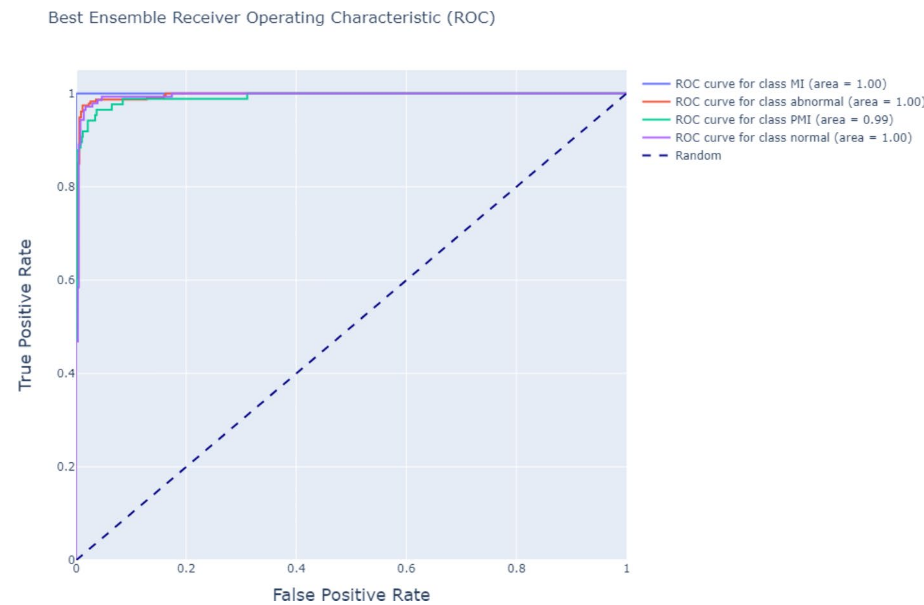

Fig. 10 The ROC curve for best ensemble model

ROC curve analysis for best ensemble model

The ROC curve for each class as explained in Fig. 10 demonstrates the ensemble's ability to differentiate each condition with a near-perfect Area AUC. Each class, MI, abnormal, HMI, and normal, presents an AUC of 1.00, 1.00, 0.99, and 1.00, respectively, indicative of the ensemble's excellent performance. These metrics signify not just the accuracy of classification but also the balanced nature of the ensemble's performance across various classes. The F1 score, being a harmonic mean of precision and recall, suggests a high level of model precision and robustness. Balanced accuracy further confirms that the ensemble performed well across all classes, not just the predominant ones. The ensemble of models, excluding the lower-performing EfficientNet, achieved a harmonious balance in classification tasks for the given ECG dataset. The analysis suggests that combining models in this ensemble effectively captures the nuances of each class in the dataset, leading to a high-performance solution for ECG image classification in cardiac patients. This ensemble is applied to serve as a reliable tool in assisting medical professionals in the diagnosis and study of cardiac conditions. The overall metrics recorded F1 score with 0.9651 and balanced accuracy with 0.9640.

Results and discussion

The present study embarked on an exploration of the capacity of deep learning models to enhance the diagnostic accuracy of ECG image analysis. This is paramount as cardiovascular diseases continue to be the leading cause of global mortality, and the ECG remains a fundamental tool in their diagnosis. Our research was propelled by critical questions addressing the adaptability of pre-trained neural networks to the medical domain, the efficacy of individual CNN architectures for ECG interpretation, and the potential of ensemble learning to optimize classification performance.

Q1 Answer: transfer learning for ECG image classification

In response to our first research question on the adaptability of pre-trained models through transfer learning, our findings were affirmative. Pre-trained models were found to be highly effective when repurposed for ECG image classification, resonating with the work of [39] who found that transfer learning could aid in medical imaging tasks. The models utilized in this study, which were originally trained on ImageNet, demonstrated the ability to discern the nuanced patterns that characterize various cardiac conditions. This validates our hypothesis that the models with image recognition tasks can carry their robust feature extraction capabilities into the medical domain.

Q2 Answer: comparative performance of CNN architectures

Our investigation into the second research question revealed a spectrum of performance across various CNN architectures. Consistent with the findings of [40], our study found that architectures like Inception and DenseNet yielded superior performance, indicating their suitability for ECG analysis. However, not all architectures displayed the same level of efficacy, underscoring the necessity of model selection based on task-specific criteria. This aligns with the notion that architectural features, such as inception modules and dense connectivity, may be particularly conducive to capturing the relevant patterns in cardiac signal data.

Q3 Answer: enhancing accuracy with ensemble learning

Addressing our third research question on the enhancement of accuracy through ensemble methods, the study confirmed that ensemble models outperformed their individual counterparts [41]. The optimal ensemble, comprising Inception, MobileNet, and NASNetLarge, achieved the highest F1 score and balanced accuracy. The incremental addition of models to the ensemble improved performance up to a threshold, beyond which the benefits plateaued. This phenomenon illustrates the law of diminishing returns in model complexity, emphasizing the need for a judicious balance between model capacity and computational efficiency.

Implications and future directions

The implications of this study are base on two main folds:

First: it provides a practical blueprint for implementing deep learning models in the analysis of ECG images, which could significantly assist cardiologists in diagnosing various cardiac anomalies more efficiently. By demonstrating that these models can be repurposed with transfer learning, we pave the way for their integration into clinical settings where they can serve as a second opinion for practitioners, thereby potentially reducing diagnostic errors.

Second: the success of ensemble models in this study suggests a new frontier in medical imaging where combinations of models are utilized rather than single architectures. The improved performance of ensemble models can be leveraged to create more reliable and accurate diagnostic tools, which is particularly important in cardiology where the cost of false negatives can be life threatening.

However, this study is not without limitations. The models were tested on a dataset from a single source, which may not fully represent the diversity of ECG patterns observed in the broader population. Future research should aim to validate these findings across multi-source datasets to ensure the models' robustness and generalizability. Additionally, the interpretability of deep learning models remains a challenge. Ensuring that these models provide insights into their decision-making processes is crucial for their acceptance by medical professionals.

Future research should also investigate the integration of ECG data with other patient information, such as demographics and clinical history, to create comprehensive predictive models. Moreover, further exploration of the cost-benefit trade-off associated with model complexity and computational resources is necessary, especially considering the deployment of such models in resource-constrained environments. In conclusion, our research highlights the potential of deep learning in the realm of cardiac health, offering a compelling case for the broader adoption of AI tools in medicine. As the field advances, a continuous dialogue between AI practitioners and clinical professionals will be essential to harness the full potential of these technologies for patient care.

Conclusion

This research represents a significant advancement in the application of deep learning techniques to the classification of electrocardiogram (ECG) images, a critical component in diagnosing cardiovascular diseases (CVDs). By systematically evaluating the efficacy of pre-trained neural networks adapted via transfer learning, our study has illuminated the substantial promise these technologies hold for enhancing diagnostic precision in cardiology. A key achievement of our work is the effective demonstration of the adaptability of pre-trained models to the domain of ECG image analysis. Models initially trained on diverse image datasets such as ImageNet have shown remarkable proficiency in identifying nuanced patterns indicative of various cardiac conditions, displaying the potential for cross-domain application of deep learning models. This adaptability underscores the versatility of deep learning in medical imaging and its capacity to significantly aid in the early detection of CVDs, potentially improving patient care outcomes. Furthermore, our comparative analysis across different convolutional neural network (CNN) architectures has yielded insightful findings. Notably, architectures like Inception and DenseNet were identified as particularly potent for ECG interpretation, underscoring the importance of architectural choices in developing AI-driven diagnostic tools. These models' superior performance highlights their potential as cornerstone technologies in the future of automated cardiac diagnosis. The exploration of ensemble learning methods stands out as a monumental achievement of this study. The construction and evaluation of various ensemble configurations have unequivocally demonstrated that a collaborative approach, wherein multiple models combine their predictive capacities, significantly enhances classification accuracy. The optimal ensemble, comprising Inception, MobileNet, and NASNetLarge, achieved unparalleled F1 scores and balanced accuracy, marking a pioneering step towards the development of highly reliable ECG classification systems. The successful implementation of these ensemble models not only sets a new benchmark in ECG image classification but also opens the door to their application in clinical settings. The study's outcomes advocate for the integration of AI in supporting diagnostic processes, offering a model for future research in medical image analysis and beyond. In summary, our investigation has achieved a critical milestone in the quest for AI-assisted diagnosis of cardiovascular diseases. Through the innovative application of deep learning models, we have explained the potential to elevate the standard of care in cardiology, enhancing the accuracy, efficiency, and reliability of cardiac disease diagnostics. This research paves the way for future advancements in the field, promising a future where technology and medicine converge to save lives and improve patient outcomes.

Acknowledgements
The authors extend their appreciation to the Deputyship of Research & Innovation, Ministry of Education in Saudi Arabia funding this research through the project number 223202.

Author contributions
Data curation, A.A.; A.M.M; N.A. and Al.A.M.; Formal analysis, S.A.; M.A.; and M.E.; Investigation, A.A.; A.M.M. and Al.A.M.; Supervision, A.A.; S.A. and M.E.; Writing—original draft, H.S.; A.A. ; Mj.Al. and A.M.M; Writing—review and editing, Mj.Al.; A.A. ; M.E.; All authors have read and agreed to the published version of the manuscript.

Funding
The Deputyship of Research & Innovation, Ministry of Education in Saudi Arabia funding this research through the project number 223202.

Data availability
ECG Images dataset of Cardiac Patients, https://data.mendeley.com/datasets/gwbz3fsgp8/2https://data.mendeley.com/datasets/gwbz3fsgp8/2 doi: 10.17632/gwbz3fsgp8.2.

Declarations

Institutional review board statement
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Conflict of interest
The authors declare no conflict of interest.

Received: 3 April 2024 / Accepted: 9 January 2025
Published online: 19 January 2025

References

1. Rechciński T. What else can AI see in a Digital ECG? J Personalized Med. 2023; 13(7).
2. Kashou AH, May AM, Noseworthy PA. Comparison of two artificial intelligence-augmented ECG approaches: machine learning and deep learning. J Electrocardiol. 2023;79:75–80.
3. Avula V, Wu KC, Carrick RT. Clinical applications, methodology, and scientific reporting of Electrocardiogram Deep-Learning models: a systematic review. JACC: Adv. 2023;2(10):100686.
4. Huang J-S, Chen B-Q, Zeng N-Y, Cao X-C, Li Y. Accurate classification of ECG arrhythmia using MOWPT enhanced fast compression deep learning networks. J Ambient Intell Humaniz Comput. 2023;14(5):5703–20.
5. Wang J. A deep learning approach for atrial fibrillation signals classification based on convolutional and modified Elman neural network. Future Generation Comput Syst. 2020;102:670–9.
6. Giriprasad Gaddam P, Sanjeeva reddy A, Sreehari RV. automatic classification of cardiac arrhythmias based on ECG signals using transferred deep learning convolution neural network. Journal of Physics: Conference Series. 2021;2089(1):012058.
7. Huang J, Chen B, Yao B, He W. ECG arrhythmia classification using STFT-Based spectrogram and convolutional neural network. IEEE Access. 2019;7:92871–80.
8. Li C, Zhao H, Lu W, Leng X, Wang L, Lin X, et al. DeepECG: image-based electrocardiogram interpretation with deep convolutional neural networks. Biomed Signal Process Control. 2021;69:102824.
9. Xiao Q, Lee K, Mokhtar SA, Ismail I, Pauzi AL, Zhang Q et al. Deep learning-based ECG arrhythmia classification: a systematic review. Appl Sci. 2023; 13(8).
10. Ahmad Z, Tabassum A, Guan L, Khan NM. ECG heartbeat classification using Multimodal Fusion. IEEE Access. 2021;9:100615–26.
11. Khan FYX, Yuan Z, Rehman, Au. ECG classification using 1-D convolutional deep residual neural network. PLoS ONE. 2023;18(4).
12. Ribeiro E, Soares FDQ, Krieger J, Gutierrez M. Deep Learning Approach for Detection of Atrial Fibrillation and Atrial Flutter based on ECG images. Anais do XXIII Simpósio Brasileiro De Computação Aplicada à Saúde, São PauloSP.509 – 14.
13. Bozyigit F, Erdemir F, Sahin M, Kilinc D, editors. Classification of electrocardiogram (ECG) data using deep learning methods. 2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT). 2020;22–24.
14. Liu X, Wang H, Li Z, Qin L. Deep learning in ECG diagnosis: a review. Knowl Based Syst. 2021;227:107187.
15. Murat F, Yildirim O, Talo M, Baloglu UB, Demir Y, Acharya UR. Application of deep learning techniques for heartbeats detection using ECG signals-analysis and review. Comput Biol Med. 2020;120:103726.
16. Ebrahimi Z, Loni M, Daneshtalab M, Gharehbaghi A. A review on deep learning methods for ECG arrhythmia classification. Expert Syst Applications: X. 2020;7:100033.

17. Ahmad Z, Tabassum A, Guan L, Khan N, editors. ECG Heart-Beat classification using multimodal image fusion. ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2021;6–11.
18. Bortolan G, editor. 3-D ECG images with Deep Learning Approach for Identification of Cardiac Abnormalities from a Variable Number of Leads. 2021 Computing in Cardiology (CinC). 2021;13–15.
19. Narotamo H, Dias M, Santos R, Carreiro AV, Gamboa H, Silveira M. Deep learning for ECG classification: a comparative study of 1D and 2D representations and multimodal fusion approaches. Biomed Signal Process Control. 2024;93:106141.
20. Got A, Zouache D, Moussaoui A, Abualigah L, Alsayat A. Improved Manta Ray foraging optimizer-based SVM for feature selection problems: A Medical Case Study. J Bionic Eng. 2024;21(1):409–25.
21. Alsayat A, Elmezain M, Alanazi S, Alruily M, Mostafa AM, Said W. Multi-layer Preprocessing and U-Net with residual attention block for retinal blood vessel segmentation. Diagnostics. 2023;13(21).
22. Almuayqil SN, Elbashir MK, Ezz M, Mohammed M, Mostafa AM, Alruily M et al. An Approach for Cancer-Type Classification Using Feature Selection Techniques with Convolutional Neural Network. Applied Sciences. 2023;13(19).
23. Mantravadi A, Saini S, Mittal RSCT, Shah S. CLINet: a novel deep learning network for ECG signal classification. J Electrocardiol. 2024;83:41–8.
24. Sadad T, Safran M, Khan I, Alfarhood S, Khan R, Ashraf I. Efficient classification of ECG images using a lightweight CNN with attention Module and IoT. Sensors. 2023; 23(18).
25. Hassaballah M, Wazery YM, Ibrahim IE, Farag A. ECG heartbeat classification using machine learning and metaheuristic optimization for Smart Healthcare systems. Bioengineering. 2023; 10(4).
26. Chui KT, Gupta BB, Zhao M, Malibari A, Arya V, Alhalabi W et al. Enhancing Electrocardiogram classification with multiple datasets and distant transfer learning. Bioeng [Internet] 2022; 9(11).
27. *ECG Images dataset of Cardiac Patients*, https://data.mendeley.com/datasets/gwbz3fsgp8/2 https://doi.org/10.17632/gwbz3fsgp8.2
28. Marocchi M, Abbott L, Rong Y, Nordholm S, Dwivedi G. Abnormal heart sound classification and model interpretability: a transfer Learning Approach with Deep Learning. J Vascular Dis. 2023; 2(4):[438 – 59 pp.].
29. Nawaz M, Saleem S, Masood M, Rashid J, Nazir T, COVID-ECG-RSNet. COVID-19 classification from ECG images using swish-based improved ResNet model. Biomed Signal Process Control. 2024;89:105801.
30. Chen W-W, Tseng C-C, Huang C-C, Lu HH-S. Improving deep-learning electrocardiogram classification with an effective coloring method. Artif Intell Med. 2024;149:102809.
31. Nurtsani AM, Zakaria H, editors. Class-Oriented and Subject-Oriented Approaches for Arrhythmia Classification Using Two-Channels ECG with MobileNet-Based CNN. 2023 International Conference on Artificial Intelligence, Blockchain, Cloud Computing, and Data Analytics (ICoABCD). 2023;13–15 Nov.
32. Ji L, Wei Z, Hao J, Wang C. An intelligent diagnostic method of ECG signal based on Markov transition field and a ResNet. Comput Methods Programs Biomed. 2023;242:107784.
33. Dessai A, Virani H. Emotion classification based on CWT of ECG and GSR signals using various CNN models. Electron [Internet]. 2023; 12(13).
34. Abo-Zahhad MM, Hassan AMA, editors classification of ECG signals for detecting coronary heart diseases using deep transfer learning techniques. 2022 10th International Japan-Africa Conference on Electronics, Communications, and Computations (JAC-ECC). 2022;19–20 Dec.
35. Mishra N, Jahan I, Nadeem MR, Sharma V, editors. A comparative study of ResNet50, EfficientNetB7, InceptionV3, VGG16 models in crop and weed classification. 2023 4th International Conference on Intelligent Engineering and Management (ICIEM). 2023;9–11 May.
36. Chorney W, Wang H. Towards federated transfer learning in electrocardiogram signal analysis. Comput Biol Med. 2024;170:107984.
37. Yoo H, Yum Y, Kim Y, Kim J-H, Park H-J, Joo HJ. Restoration of missing or low-quality 12-lead ECG signals using ensemble deep-learning model with optimal combination. Biomed Signal Process Control. 2023;83:104690.
38. Alsayat A, Ahmadi H. A hybrid method using ensembles of neural network and Text Mining for Learner Satisfaction Analysis from Big datasets in Online Learning platform. Neural Process Lett. 2023;55(3):3267–303.
39. Yu X, Wang J, Hong Q-Q, Teku R, Wang S-H, Zhang Y-D. Transfer learning for medical images analyses: a survey. Neurocomputing. 2022;489:230–54.
40. Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med. 2019;25(1):65–9.
41. Roy M, Baruah U. Enhancing medical image classification through controlled diversity in ensemble learning. Eng Appl Artif Intell. 2024;133:108138.

Publisher's note