xECG-Beats: an explainable deep transfer learning approach for ECG-based heartbeat classification

Peimankar, Abdolrahman; Ebrahimi, Ali; Wiil, Uffe Kock

Published in:
Network Modeling Analysis in Health Informatics and Bioinformatics

DOI:
10.1007/s13721-024-00481-2

Publication date:
2024

Document version:
Final published version

Document license:
CC BY

Citation for pulished version (APA):
Peimankar, A., Ebrahimi, A., & Wiil, U. K. (2024). xECG-Beats: an explainable deep transfer learning approach for ECG-based heartbeat classification. *Network Modeling Analysis in Health Informatics and Bioinformatics*, *13*(1), Article 45. https://doi.org/10.1007/s13721-024-00481-2

Network Modeling Analysis in Health Informatics and Bioinformatics (2024) 13:45
https://doi.org/10.1007/s13721-024-00481-2

ORIGINAL ARTICLE

xECG-Beats: an explainable deep transfer learning approach for ECG-based heartbeat classification

Abdolrahman Peimankar[1] · Ali Ebrahimi[1] · Uffe Kock Wiil[1]

Received: 10 November 2023 / Revised: 19 June 2024 / Accepted: 9 August 2024

Abstract
Early detection of abnormal heartbeats is of great importance for cardiologists for early diagnosis of cardiac diseases. This will help patients to receive in time diagnosis and prevention. Conventionally, physicians provide cardiac diagnoses by visual examination of electrocardiograms (ECGs). However, this can be a very time consuming and demanding task and, in some cases, may lead to overlooking and wrong diagnosis of life-threatening heart diseases. Therefore, an intelligent model can help to automatically analyze these huge amount of ECGs captured by different devices in clinical practice. A deep transfer learning approach is used to utilize the capability of different trained deep neural networks and to test them on new unseen datasets without the need to fully re-train the model. Two deep neural networks, namely, Visual Geometry Group (VGG) and Residual Network (ResNet) are utilized for classification of ECGs heartbeats. The models are evaluated using two unseen ECG datasets (i.e., SVDB and INCARTDB) by only optimizing their last classification layers. The overall area under curve for receiver operating characteristic (AUCROC) of two VGG and ResNet models are 0.961 and 0.966 on the SVDB dataset, respectively, and both models achieve 0.981 on the INCARTDB. This paper proposes an accurate and explainable model to classify ECG heartbeats into five categories recommended by the ANSI/AAMI standard. The proposed method paves the way to use pre-trained deep neural networks in real-time monitoring of heart patients using ECG data and to help clinicians understand the decision made by the models on each case using an explainable approach.

Keywords Electrocardiogram (ECG) · Heartbeat classification · Deep learning · Transfer learning · Explainable AI

1 Introduction

Cardiac arrhythmia is an important issue due to its prevalence, high mortality rates, and the considerable expenses involved in its treatment (De Chazal and Reilly 2006; Yıldırım et al. 2018; Mondéjar-Guerra et al. 2019). The World Health Organization (WHO) identifies cardiovascular diseases as a leading global cause of death (World Health Organization 2023).

Electrocardiography (ECG) is the most basic and widely available technique to diagnose cardiac arrhythmias, which provides useful information for the healthcare providers on the patients' cardiovascular conditions (Yıldırım et al. 2018; Faezipour et al. 2010; Das and Ari 2014). However, visual examination of these long ECG recordings, in most cases, can be a very challenging and time consuming task for the physicians, which may take up to several hours or days for some specific arrhythmias (De Chazal and Reilly 2006; Faezipour et al. 2010). Thus, intelligent algorithms along with advanced computational techniques are needed for automatic analysis of ECGs, which can be of a great assistance to healthcare providers and significantly release available resources for better treatment of cardiac diseases (De Chazal and Reilly 2006; Faezipour et al. 2010; Das and Ari 2014; Luz et al. 2016).

Various state-of-the-art Machine Learning (ML) algorithms have been proposed for automatic classification and detection of heartbeats using ECGs. As an example, De Chazal and Reilly (2006) proposed a linear discriminant

✉ Abdolrahman Peimankar
abpe@mmmi.sdu.dk

Ali Ebrahimi
aleb@mmmi.sdu.dk

Uffe Kock Wiil
ukwiil@mmmi.sdu.dk

[1] SDU Health Informatics and Technology, The Maersk Mc-Kinney Moller Institute, University of Southern Denmark, 5230 Odense, Denmark

Published online: 23 August 2024

analysis classifier using three types of features extracted from ECG signals for the classification of heartbeats into five different types (Table 1). Faezipour et al. (2010), applied a wavelet based approach for automatic classification of heartbeats. Although, their proposed model achieved relatively high performance, an extensive data preprocessing is required and the proposed method was not tested on external new test sets. In another study, Das and Ari (2014) developed a neural network model for ECG beats classification using an extensive feature extraction step, which makes such models not suitable for implementation in clinical settings.

Many studies have also used support vector machine (SVM) to diagnose heart arrhythmias using ECG signals (Mondéjar-Guerra et al. 2019; Melgani and Bazi 2008; Ye et al. 2012). For example, Mondéjar-Guerra et al. (2019) developed an ensemble of SVMs for the heartbeat classification using both temporal and morphological features of ECG recordings. Their proposed ensemble SVM model shows significant improvement over a single SVM classifier. However, the model has been only validated on the same training dataset. In addition, Melgani and Bazi (2008) compared the performance of SVM algorithm for classifying ECG beats with two other algorithms namely k-Nearest Neighbors (kNN) and Radial Basis Function (RBF) based neural network. They have applied a Particle Swarm Optimization (PSO) algorithm to find the optimum parameters of SVM, which resulted in higher predictive performance. The main drawback of their proposed PSO-SVM is the computational complexity of PSO algorithm, which can be very time consuming in practice on the new datasets and needs many parameters to be set.

Furthermore, there are many other studies that focus on the application of Deep Learning (DL) for ECG signal analysis and cardiac diseases diagnosis (Alonso-Atienza et al. 2012; Peimankar and Puthusserypady 2021; Sannino and De Pietro 2018; Mousavi and Afghah 2019; Andersen et al. 2019; Murat et al. 2020; Sellami and Hwang 2019; Jahan et al. 2022; Peimankar and Puthusserypady 2019, 2018). Despite significant progress in ML and DL algorithms for ECG analysis, there remains a gap in understanding their applicability across diverse patient populations and the interpretability of their decisions. For example, Alonso-Atienza et al. (2012) proposes a novel feature selection algorithm for early detection of ventricular fibrillation in ECG signals, combining SVM with bootstrap resampling. This method improves detection efficiency by reducing the feature set while maintaining performance. The algorithm was tested on two different databases, showing superior performance compared to existing methods. However, there are some drawbacks with this algorithms such as high computational burden due to the bootstrap resampling process, dependency on tuning the parameters for optimal performance, and lack of explainability analysis. Sannino and De Pietro (2018) presents a deep neural network approach for ECG beat classification. The model, tested on a single database, demonstrates superior accuracy, sensitivity, and specificity compared to existing methods and its potential for real-time application in clinical settings. But, the drawbacks of their proposed method include a dependency on a substantial amount of annotated data for training, the need for further validation using additional datasets, and high computational demands, which may complicate real-time applications. Murat et al. (2020) reviews and evaluates deep learning methods for ECG arrhythmia detection, highlighting various models and experimental studies. It utilized a five-class ECG dataset with 100,022 beats to analyze the performance of different deep learning techniques. The study emphasizes the advantages of deep learning in handling raw ECG signals without manual feature extraction and discusses the challenges such as dataset imbalance and model optimization. The primary drawbacks of using deep learning for ECG arrhythmia detection include high computational costs, as these models require significant resources. Sellami and Hwang (2019) presents a novel deep convolutional neural network for accurate heartbeat classification using raw ECG signals without preprocessing. This approach addresses class imbalance with a dynamic batch-weighted loss function and uses multiple heartbeats for better classification. It achieves high performance metrics, significantly outperforming existing methods. This paper identifies a few drawbacks in its proposed method for heartbeat classification. For example, it requires large, annotated datasets of heartbeats, which are costly and time-consuming to obtain as they need to be labeled by clinical experts. This limitation can hinder the model's applicability and scalability.

Key questions persist regarding the generalizability of DL models trained on one cohort to entirely different patient demographics and how to ensure clinicians can interpret complex DL model outputs effectively. To address these questions, in this study, we followed the below objectives:

- Investigate the feasibility and effectiveness of transfer learning using modified DL architectures for analyzing 1-D time series ECG signals, a departure from conventional 2-D image-based approaches.
- Assess the transferability of DL models trained on one dataset to entirely unseen datasets, evaluating their performance on different patient cohorts. It should be noted that none of the above mentioned studies has applied transfer learning. We test the trained DL model on two completely unseen test sets by freezing the weights of the trained model. We only use 5% of the test sets to train the final classification layers and the rest (95%) for evaluating the new datasets, which makes it possible to implement such large models in clinical settings where there is not enough data to retrain such models.

- Develop and evaluate explainable AI techniques to explain DL model decisions, enabling healthcare professionals to comprehend and trust the automated diagnostic outputs.
- Implement and validate two distinct transfer learning models (VGG and ResNet) using publicly available ECG databases, assessing their performance and generalizability across diverse patient populations.
- Compare the efficacy of transfer learning-based DL models with traditional ML algorithms and explore their potential for real-world clinical applications.

We develop and evaluate two different transfer learning models (i.e., VGG and ResNet) using three publicly available ECG databases. Transfer learning methods are being used in many different domains such as health informatics (Mohammad and Saeed 2022; Zhu et al. 2020; Gaur et al. 2022; Cheplygina et al. 2017; Ahsan et al. 2023), energy (Himeur et al. 2022; Liu et al. 2021; Fan et al. 2020; Hua et al. 2022; González-Vidal et al. 2022), and finance (Li et al. 2018; Wu et al. 2022; Yu et al. 2018; Chen et al. 2022). It has been widely shown in the literature that transfer learning helps generalizing DL models. Therefore, in most cases transfer learning is favourable for utilizing DL models in real-world applications.

The rest of this paper consists of three sections. Section 2 presents the proposed method used in this study in addition to the description of the datasets, pre-processing, post-processing, and models applied in this paper. The experimental results of the proposed models are presented and discussed in Sect. 3. Lastly, Sect. 4 provides the conclusion of this study.

2 Materials and methods

2.1 Databases

In this study, the publicly available MIT-BIH Arrhythmias Database (MITDB) is used to train the DL models (Moody and Mark 2001; Goldberger et al. 2000). The MITDB contains 48 ECG recordings each 30 min long and sampled at 360 Hz. The recorded ECGs of 17 subjects only contain "Normal" rhythms. The amplitude of raw ECG signals ranges between ±10mV that have been converted into digital signals of 11-bit resolution. Following the AAMI recommendation, four records/subjects (102, 104, 107, and 217) were excluded from the database since they contain paced beats without appropriate signal quality. The MITDB also provides beat annotation along with the raw ECG signals. As given in Table 1, the ANSI/AAMI standard is utilized to categorize the beats' annotations into five classes (Testing and Reporting 2023). Figure 1 shows heartbeats examples of the five classes listed in Table 1. The MIT-BIH Supraventricular Arrhythmia Database (SVDB) and the St Petersburg INCART Arrhythmia Database (INCARTDB) are used to examine the DL models on completely unseen data. The SVDB has 78 ECG recordings of length 30 min sampled at 360 Hz (Greenwald et al. 1990). Most of the recordings contains supraventricular arrhythmias. The INCARTDB includes 75 ECG recordings that are extracted from 32 Holter monitors (17 male and 15 female patients), which are half-hour long each and sampled at 257 Hz (Tihonenko et al. 2008). Most of these patients had ventricular beats and suffered from ischemia, coronary artery disease, conduction abnormalities, and arrhythmias. A summary of these three databases is provided in Table 2.

2.2 Data preprocessing

To prepare ECG signals for input into the DL models, several preprocessing steps have been applied. These steps ensure consistency and compatibility across different databases and facilitate accurate analysis by the models.

1. First, we esampled the INCARTDB to match the 360 Hz frequency of the MITDB and SVDB databases.

Table 1 AAMI recommendation for mapping heartbeats categories

Category	Heartbeats mapped classes
N	Normal beat
	Left and right bundle branch block beats
	Atrial escape beat
	Nodal escape beat
S	Atrial premature beat
	Aberrated atrial premature beat
	Supra ventricular premature beat
	Nodal premature beat
V	Premature Ventricular contraction beat
	Ventricular escape beat
F	Fusion of normal and ventricular beat
Q	Paced beat
	Fusion of paced and normal Beat
	Unclassified beat

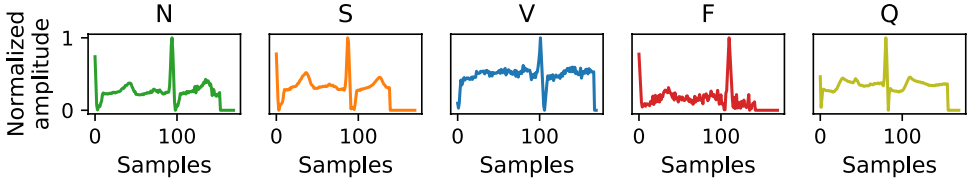


Fig. 1 An example of five different heartbeats categories/classes mapped using AAMI standard

Table 2 Total number of extracted heartbeats for the three databases

Dataset	# of hearteats
MITDB	65,957
SVDB	127,007
INCARTDB	120,042

2. The long ECG recordings were also segmented into 5-second intervals to normalize their amplitude between zero and one (see Fig. 1).
3. Next, the location of R-peaks were determined in the segmented ECG signals using the provided annotations from the databases.
4. Subsequently, we calculated the R-R distances between the current heartbeat and the two adjacent beats.
5. Then, single heartbeats were extract with a length equal to the average of the two adjacent R-R intervals.
6. Finally, the extracted heartbeats were zero padded (see Fig. 1) to have the same length as required for the inputs of the DL models.

2.3 Deep learning models

DL models have revolutionized different fields by enabling automated learning and feature extraction from input data. Time series classification is a challenging task in various domains. While Convolutional Neural Networks (CNNs) have proven to be highly effective in image classification, their application to time series data requires adaptations to account for the temporal nature of the data. In this section, we present an overview of adapting the VGG and ResNet models, originally designed for image classification, to the task of time series classification (Ismail Fawaz et al. 2019). We discuss the architectural modifications and the key considerations in training and evaluating the two DL models for time series data. The proposed models were implemented in Python version 3.10.12, utilizing the Keras version 2.11.0 API. The Keras API stands out as a high-level neural networks API, specifically designed to facilitate rapid experimentation (Chollet 2015). Furthermore, the matrix analysis was performed using TensorFlow version 2.11.0, leveraging GPU support for enhanced computational efficiency.

2.3.1 VGG model

The VGG-16 model, proposed by Simonyan and Zisserman (2014), has emerged as a fundamental and influential architecture that has significantly advanced the state-of-the-art in classification tasks. The VGG-16 model's success lies in its ability to capture patterns by employing a deep stack of convolutional layers and exploiting large receptive fields (Simonyan and Zisserman 2014). The primary distinguishing factor of the VGG-16 model is its depth. With 16 weight

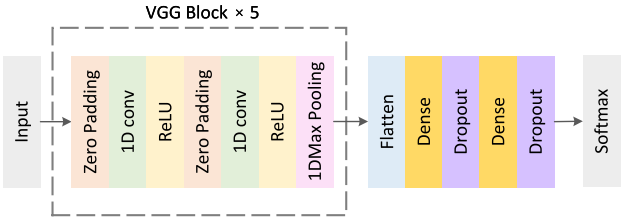


Fig. 2 VGG model architecture

layers, it possesses a considerable number of parameters, approximately 138 million, making it more complex than previous DL models. The increased depth of VGG-16 allows for a more nuanced representation of input data, capturing both low-level and high-level features. In this paper, a modified version of VGG-16 model has been used to be able to handle 1-D time series data instead of traditional image data. The architecture of the modified VGG model is depicted in Fig. 2. In total, the model has 6,448,421 trainable parameters. The model was trained using the Adaptive Moment Estimation (Adam) optimizer (Kingma and Ba 2014) and sparse categorical crossentropy as loss function with learning rate and decay of of 0.01 and 0.005, respectively. Furthermore, an early stopping technique is employed to avoid overfitting of the model. The training process is terminated if there is no reduction in the validation loss for seven consecutive epochs. Figure 3 illustrates the training curves, which show the loss and accuracy of train and validation process vs. number of epochs.

2.3.2 ResNet model

The ResNet model, proposed by He et al. (2016), has emerged as another groundbreaking architecture that addresses the challenge of training very deep neural networks. The ability to train deep neural networks with tens or hundreds of layers has been a longstanding challenge in the field of DL. The ResNet model introduced a novel residual learning framework that enabled the successful training of extremely deep neural networks by mitigating the degradation problem. By introducing shortcut connections, or skip connections, ResNet alleviates the vanishing gradient problem and facilitates the learning of deeper and more accurate representations (He et al. 2016). The core idea behind the ResNet model is residual learning, which involves learning residual mappings instead of directly learning the underlying desired mappings. Residual learning is achieved by introducing skip connections that bypass one or more layers in the network, allowing the network to learn the residual between the input and the desired output. These skip connections enable the model to propagate information more effectively throughout the network and alleviate the degradation problem, where deeper networks tend to have higher

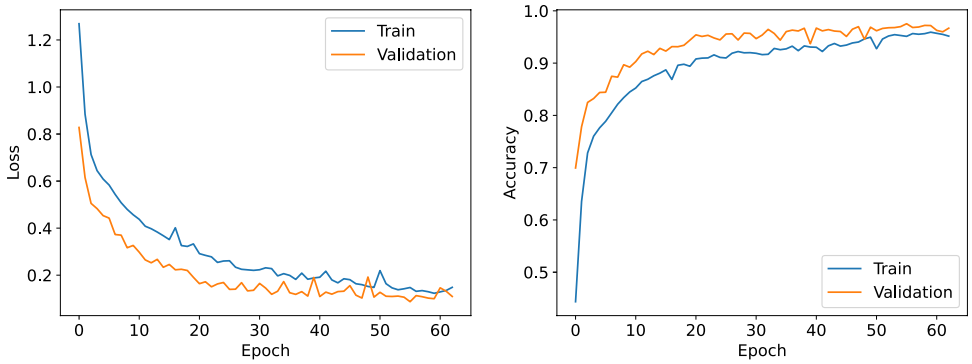


Fig. 3 Training and validation curves of VGGs model

(a) Loss (b) Accuracy

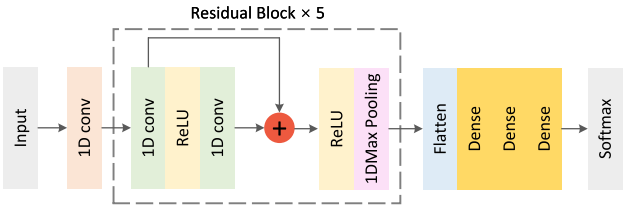


Fig. 4 ResNet model architecture

training errors (He et al. 2016). In this paper, a modified version of ResNet model has been used to be able to handle 1-D signals. The architecture of the modified ResNet model is shown in Fig. 4. In total, the model has 55,013 trainable parameters. The model was trained using the Adaptive Moment Estimation (Adam) optimizer (Kingma and Ba 2014) and sparse categorical crossentropy as loss function with learning rate and decay of of 0.01 and 0.005, respectively. Moreover, an early stopping technique is employed to avoid overfitting of the model. The training process is terminated if there is no reduction in the validation loss for seven consecutive epochs. Figure 5 illustrates the training curves, which show the loss and accuracy of train and validation process vs. number of epochs.

2.4 Explainable deep learning

Despite the remarkable performance achieved by DL models in various domains, understanding and interpreting their decisions remain challenging, particularly in tasks involving the analysis of one-dimensional (1-D) signals.

In this paper, we delve into the application of Gradient-weighted Class Activation Mapping (GRAD-CAM) (Selvaraju et al. 2017). Initially proposed for visual interpretation in image classification tasks, GRAD-CAM presents a promising avenue for enhancing the explainability of DL models (Selvaraju et al. 2017). One of the key advantages of GRAD-CAM lies in its adaptability to capture important features within 1-D signals, enabling clinicians to glean insights into the model's decision-making process and bolster interpretability.

GRAD-CAM operates as an interpretation technique leveraging the gradients of the target class score concerning the convolutional feature maps to localize significant regions within input data (Selvaraju et al. 2017). By harnessing the gradient information flowing into the final convolutional layer, GRAD-CAM generates a heat map that delineates discriminative regions pertinent to the predicted class (Selvaraju et al. 2017). This heat map acts as a visual explanation,

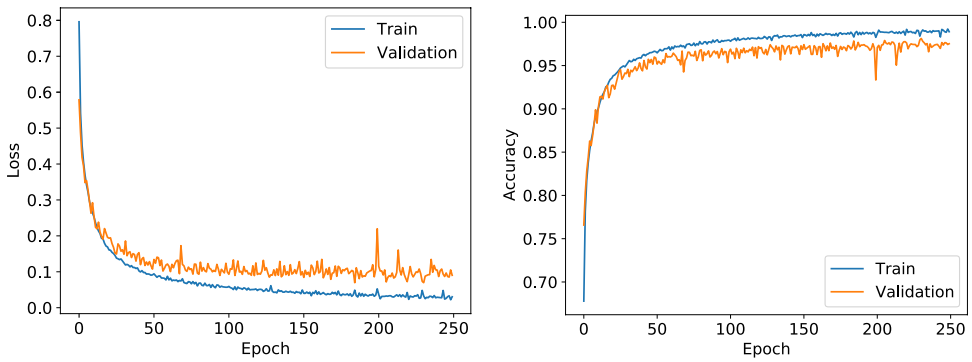


Fig. 5 Training and validation curves of ResNet model

(a) Loss (b) Accuracy

furnishing insights into the model's attentional focus and decision-making rationale.

However, adapting Grad-CAM to 1-D signals such as ECG also presents some challenges. Unlike images, which have spatial dimensions, 1-D signals are characterized by their temporal or sequential nature. As a result, interpreting the relevance of specific segments in the signal may be more complex, requiring careful consideration of the signal's temporal dynamics.

Despite these challenges, the application of GRAD-CAM to 1-D signal analysis offers significant advantages. It provides clinicians with a transparent and intuitive means of understanding DL model decisions, thereby enhancing trust and facilitating informed decision-making in various domains, including healthcare, speech processing, and beyond. Through the utilization of GRAD-CAM, we aim to elucidate the black box nature of DL models in 1-D signal analysis, ultimately promoting greater transparency and trustworthiness in their application.

2.5 Deep transfer learning based heartbeats classification framework

Figure 6 illustrates the flowchart of the proposed deep transfer learning model for heartbeat classification, which is described step-by-step as follows:

1. *Segmentation and scaling*: as mentioned in Sect. 2.2, the long ECG recordings are segmented into smaller episodes of 5 s. Then, the 5 s segments are scaled into zero mean and unit standard deviation.
2. *Heartbeats extraction*: the single heartbeats are extracted using the R-R distances between the location of the current and adjacent R-peaks (Sect. 2.2) and saved along with their corresponding annotations.
3. *Synthetic data oversampling*: As given in Table 3, the number of heartbeats for the "normal" heartbeats is much higher than the other four classes in the MITDB dataset. Therefore, the training dataset (MITDB) is very imbalanced and the number of, especially, class *V* and *Q* are not sufficient to efficiently train the DL models. Imbalanced datasets may lead to a biased classification, which consequently increases the error rate for the minority classes (i.e., *F* and *Q*) (He et al. 2008). To address this problem, the adaptive synthetic over-sampling technique (ADASYN) is used to add more heartbeats to the minority class so that the DL classifiers are able to achieve a higher and a more stable performance (He et al. 2008). This also helps improving the generalizability of the trained models on the unsenn test sets. Three main steps of ADASYN algorithms are as follows: (1) find the degree of the class imbalance in order to calculate the required synthetic heartbeats to be generated for the minority classes; (2) find the k closest heartbeats of the minority classes using Euclidean distance; and (3) generate the synthetic heartbeats for the minority classes as below:

$$d_i = x_i + (x_{ki} - x_i) \times \lambda, \tag{1}$$

where x_i represents an arbitrary sample from the minority class, x_{ki} is one of the nearest neighbor heartbeats, and λ is a random values in the range of [0, 1]. As presented in Fig. 6, data balancing (ADASYN) is only applied to the train set (MITDB) and not to the unseen test sets, which helps a realistic evaluation on the test sets.
4. *Classifiers training*: the preprocessed ECG heartbeats from MITDB is used to train the two VGG and ResNet classifiers.
5. *Transfer learning*: in this step, only 5% of the two unseen test sets (i.e., SVDB and INCARTDB) are used to reatrained the last classification layers of the DL models. All the other wights and parameters of the models remains fixed except the last classification ("Dense") layers.
6. *Classifier evaluation*: the performances of the two DL models are evaluated to examine and compare their classification capabilities. It should be noted that the models use all the test sets for performance evaluation.

Table 3 Number of heartbeats the for imbalanced and balanced datasets

Class	Imabalanced	Balanced
N	113,639	113,639
S	7946	113,847
V	5348	112,684
F	5	113,640
Q	69	113,656

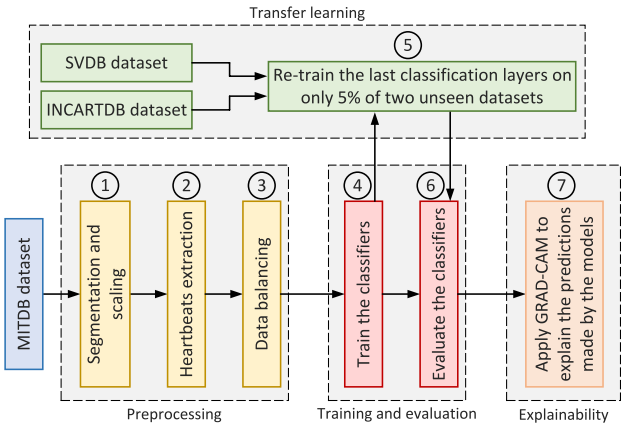


Fig. 6 Flowchart of the proposed approach. The numbers correspond to the steps in Sect. 2.5

7. *Explainability*: the predictions of the models are analyzed using GRAD-CAM technique in order to explain the decision made by the models.

3 Experimental validation

As mentioned in Sect. 2.1, the MITDB dataset (Moody and Mark 2001; Goldberger et al. 2000) was utilized to train the developed DL models. In addition, the SVDB (Greenwald et al. 1990) and INCARTDB (Tihonenko et al. 2008) datasets were used to test the DL models on new unseen test sets. The second channel of ECG recordings from all datasets have been used for the analyses.

3.1 Evaluation metrics

We use different evaluation metrics to compare the performance of the models. Table 4 presents a confusion matrix, which is used as the basis for calculating various evaluation metrics. In Table 4, the rows correspond to the actual labels, while the columns represent the predictions generated by the models. In this paper, we employ four classification metrics called accuracy (*Acc*), sensitivity (*Se*), precision (positive predictive value (*PPV*), and *F*-score. The definitions of these metrics are provided below by using the terms from Table 4:

Table 4 Confusion matrix

	Predicted negative	Predicted positive
Actual negative	True negative (TN)	False positive (FP)
Actual positive	False negative (FN)	True positive (TP)

$$Acc = \frac{TP + TN}{TP + FN + FP + TN}, \tag{2}$$

$$Se = \frac{TP}{TP + FN}, \tag{3}$$

$$PPV = \frac{FP}{FP + TN}, \tag{4}$$

$$F\text{-}score = (1 + \beta)\frac{PPV \cdot Se}{(\beta^2 \cdot PPV) + Se}. \tag{5}$$

where *TP*, *TN*, *FP*, and *FN* are the number of true positives, true negatives, false positives, and false negatives, respectively. The *F*-score is a weighted harmonic mean of *Se* and *PPV*. When the value of β is set to 1, it is referred to as the balanced *F*-score or F_1-score. The F_1-score considers both sensitivity and precision equally in its calculation.

3.2 Heartbeats classification performance on MITDB

As mentioned, MITDB was used for training the DL models. The train and validation performances of the VGG and ResNet models on the MITDB dataset for five classes are reported in Table 5. In general, the ResNet model achieves higher performance compared to VGG model taking into account both train and validation results. In addition, as shown in Table 5, both models perfectly classify class F and Q with very high classification performance, which may be due to the high oversampling of these two classes (Table 3). Although, detecting "Normal" heartbeats (Class N) is the most challenging for both VGG and ResNet models, they are classified successfully with relatively high performance.

Table 5 Comparison of classification performance on the MITDB dataset between the two DL algorithms

Algorithm	Class	Train				Validation			
		Se	PPV	F_1-score	Acc	Se	PPV	F_1-score	Acc
VGG	N	0.92	0.99	0.95	0.86	0.87	0.98	0.93	0.88
	S	1.00	0.98	0.99	0.98	1.00	0.97	0.98	0.99
	V	0.98	0.94	0.96	0.95	0.97	0.92	0.94	0.97
	F	1.00	0.99	0.99	0.98	1.00	0.98	0.99	0.99
	Q	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Avg	0.98	0.98	0.98	0.95	0.97	0.97	0.97	0.97
ResNet	N	0.96	0.99	0.98	0.94	0.91	0.98	0.94	0.93
	S	1.00	0.99	0.99	0.99	0.99	0.98	0.99	0.99
	V	0.99	0.97	0.98	0.99	0.98	0.94	0.96	0.98
	F	1.00	1.00	1.00	1.00	1.00	0.99	0.99	1.00
	Q	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Avg	0.99	0.99	0.99	0.98	0.98	0.98	0.98	0.98

The performances on the five heartbeat categories are given for the both VGG and ResNet models. The average (Avg.) of the five classes is also reported

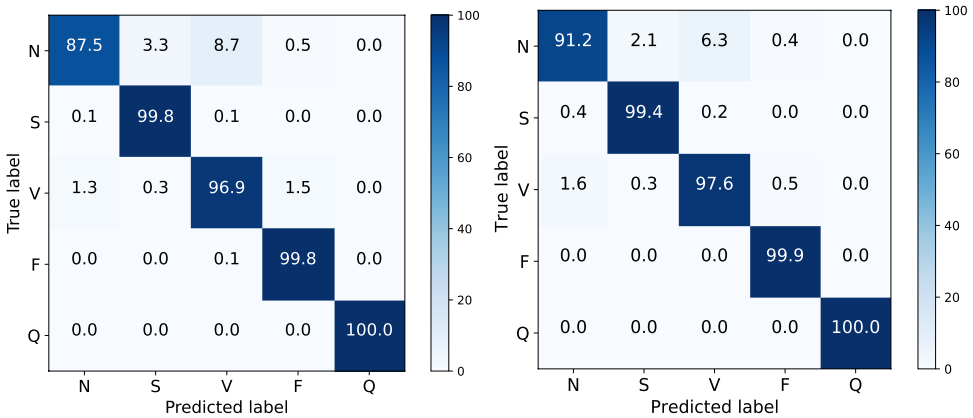
The confusion matrices of VGG and ResNet models on the validation set of the MITDB are shown in Fig. 7. Overall, the ResNet model outperforms the VGG model in terms of percentage of *TP* cases. For instance, when classifying Normal heartbeats, the VGG model achieved a *TP* rate of 87.5%, whereas the ResNet model achieved a higher rate of 91.2%. In addition, the ResNet model improves the classification performance of Ventricular (V) heartbeats by around 0.8% compared to VGG model.

The receiver operating characteristics (ROC) curves (solid lines) for five classes are depicted in Fig. 8 for both the VGG and ResNet models. Additionally, the micro- and macro-average ROC curves are shown as dashed lines (Baeza-Yates and Ribeiro-Neto 1999). The corresponding area under the curve (AUC) for each heartbeat categoryare also reported in Fig. 8. It is important to highlight that a higher AUC indicates better classification performance for the model. An AUC value of 1 represents a perfect classification performance. It can be seen from Fig. 8 that the AUCs for almost all of the classes are generally higher than 0.99. These results demonstrate the impressive capability of the two DL classifiers in effectively distinguishing between various heartbeats classes.

3.3 Model evaluation on SVDB and INCARTDB as test sets

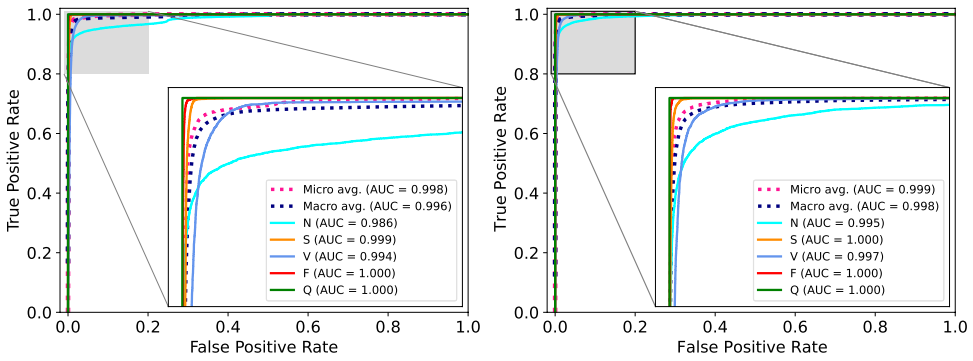
The performance of the VGG and ResNet models on the two SVDB and INCARTDB test sets are given in Table 6. As already mentioned, only the last classification layers of the two models are trained using 5% of the test datasets. Thus, the results reported in Table 6 are on the 95% of the two test datasets. Although the ResNet model performs better on the validation results of the MITDB dataset (Table 5), the VGG model generalizes better on the test sets.

The confusion matrices of the VGG and ResNet models for the two unseen test sets (i.e., SVDB and INCARTDB) are shown in Fig. 9. It can be seen from Fig. 9 that the VGG model performs better than the ResNet model on the S and V classes. On the other hand, the ResNet model completely outperforms the VGG model for classification on Normal heartbeats. For example, the ResNet model increases the detection of *TP* cases by around 6 and 10% for SVDB and INCARTDB datasets, respectively. From Fig. 9, most of the incorrect classification cases of Normal heartbeats are detected as class S and V.



(a) VGG (b) ResNet

Fig. 7 Confusion matrices of VGG and ResNet models on the validation set of the MITDB. The numbers are in percentage



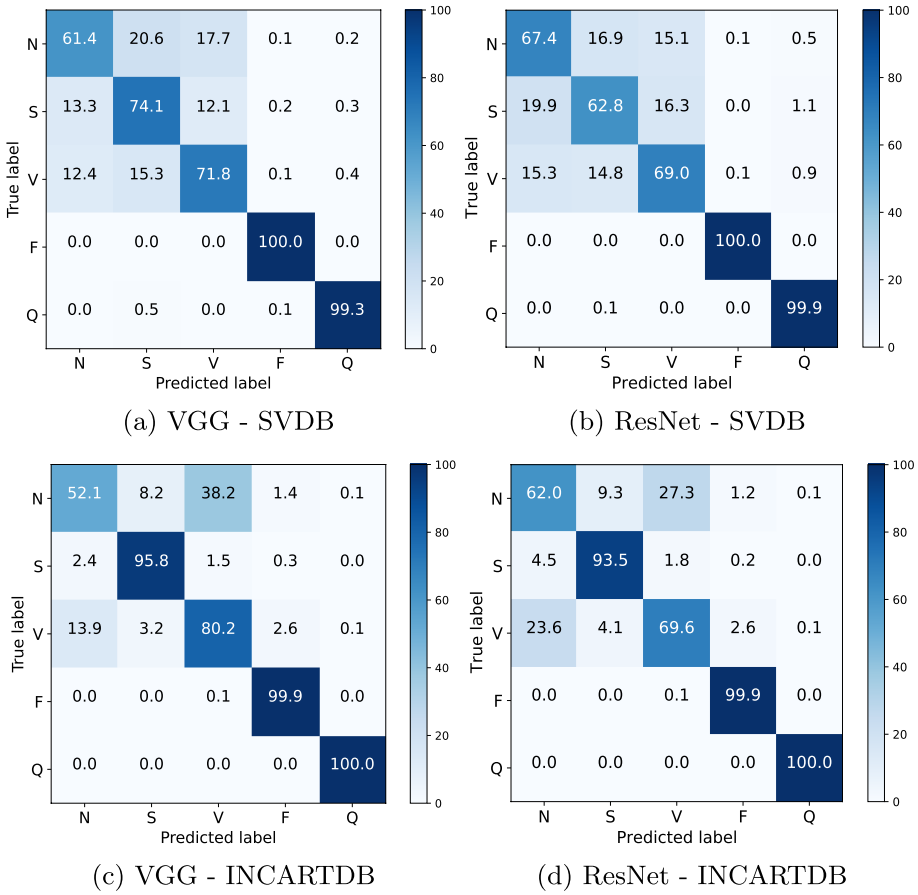
(a) VGG (b) ResNet

Fig. 8 ROC curves (solid lines) for five classes using both the VGG and ResNet models along with the the micro- and macro-average ROC curves (dashed lines) on the validation set of the MITDB

Table 6 Comparison of classification performance on the SVDB and INCARTDB datasets between the two DL algorithms

Algorithm	Class	SVDB				INCARTDB			
		Se	PPV	F_1-score	Acc	Se	PPV	F_1-score	Acc
VGG	N	0.61	0.71	0.66	0.61	0.52	0.76	0.62	0.52
	S	0.74	0.67	0.70	0.74	0.96	0.89	0.92	0.96
	V	0.72	0.70	0.71	0.72	0.80	0.67	0.73	0.80
	F	1.00	1.00	1.00	1.00	1.00	0.96	0.98	1.00
	Q	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00
	Avg	0.81	0.81	0.81	0.81	0.86	0.86	0.85	0.86
ResNet	N	0.67	0.66	0.67	0.67	0.62	0.69	0.65	0.62
	S	0.63	0.67	0.65	0.63	0.93	0.87	0.90	0.94
	V	0.69	0.69	0.69	0.69	0.70	0.70	0.70	0.70
	F	1.00	1.00	1.00	1.00	1.00	0.96	0.98	1.00
	Q	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Avg	0.80	0.80	0.80	0.80	0.85	0.85	0.85	0.86

The performances on the five heartbeats categories are given for the both VGG and ResNet models. The average (Avg.) of the five classes is also reported

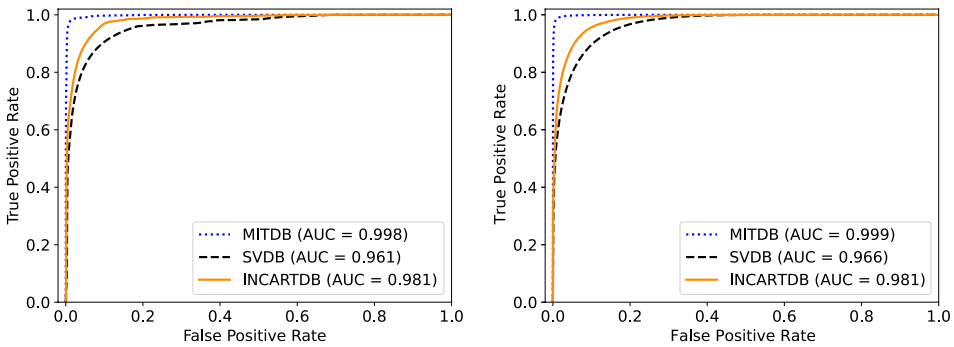


(a) VGG - SVDB (b) ResNet - SVDB

(c) VGG - INCARTDB (d) ResNet - INCARTDB

Fig. 9 Confusion matrices of VGG and ResNet models for the SVDB and INCARTDB as unseen test sets. The numbers are in percentage

As shown in Fig. 10, the performance of the two DL models on both the train set (MITDB) and test sets (SVDB and INCARTDB) are compared using overall ROC curves. Both models generalize and perform better on the INCARTDB with AUC-ROC of 0.981. The performance on the SVDB



(a) VGG (b) ResNet

Fig. 10 Comparison of average ROC curves of the VGG and ResNet algorithms on the MITDB, SVDB, and INCARTDB datasets.The area under the curves for each datasets are also given

dataset is also relatively high taking into account the AUC-ROC of 0.961.

3.4 Comparison of xECG-beats model with other state-of-the-art methods

Table 7 presents a selection of state-of-the-art models, considered by us, in the literature, all of which have utilized the MIT-BIH database to evaluate their algorithms. Our proposed classifier displayed superior performance compared to most of the models listed in Table 7. Although, the results presented in Lal et al. (2023), Kumar et al. (2023); Kallas et al. (2012), and (Engin 2004) are comparable with our proposed VGG and ResNet models, most of them (except the first study) either have considered fewer number of classes (e.g., 3 and 4 types of heartbeats) or an extensive feature engineering was done to prepare the inputs for their models. For example, Engin (2004) have used fewer number of classes and various feature types have been extracted from the raw ECG signals, which makes it very challenging to deploy such models in clinical settings due to the potential lack of flexibility and generalizability on the new datasets.

3.5 Explainability analysis

Figure 11 visualizes examples of *TP* and *FN* cases for both DL models using GRAD-CAM method. The darker areas in the figure represents the locations that the models pay more attention to in their decision making process. As shown in Fig. 11a and b, the models place more weight and attention on the T-wave appearance, which is one of the characteristics of the premature ventricular contraction beats. Moreover, the premature ventricular contraction beats occurs earlier than the normal beats. Subsequently, in 11b, the ResNet model explores the time interval between two adjacent beats as well. On the other hand, for the *FN* cases (Fig. 11c and d) mostly looks at the pre-QRS complexes and P-wave, which does not help the detection of V class (Premature Ventricular Contraction and Ventricular Escape beats).

As shown in this section, explainable AI and interpretability approaches hold significant promise in aiding

Table 7 Comparison of heartbeats classification on MITDB dataset between xECG-Beats and other state-of-the-art methods

References	Classifier	# of classes	Feature	Acc (%)
Lal et al. (2023)	CNN	5	Raw ECG	98
Kumar et al. (2023)	Fuzzy clustering and DNN	5	Raw ECG	97
Sharma et al. (2021)	NNs	5	Morphological, RR-interval	96
Shi et al. (2019)	Ensemble of classifiers	5	Region feature extraction	75
Guo et al. (2019)	CNN, Gated recurrent unit	5	Raw ECG	94
Peimankar et al. (2019)	Ensemble of classifiers	5	RR-interval, morphological, wavelets	96
Lassoued and Ketata (2018)	NNs	5	Morphological, wavelet	94
Kallas et al. (2012)	SVM	3	Kernel PCA	97
Güler and Übeylı (2005)	Ensemble of neural networks	4	Wavelet	96
Engin (2004)	Minimum distance, kNN, Bayes	4	Higher-order statistics, wavelet	98
Chen et al. (1996)	Set of rules	2	RR-interval	95
This work	VGG	5	Raw ECG	97
	ResNet	5	Raw ECG	98

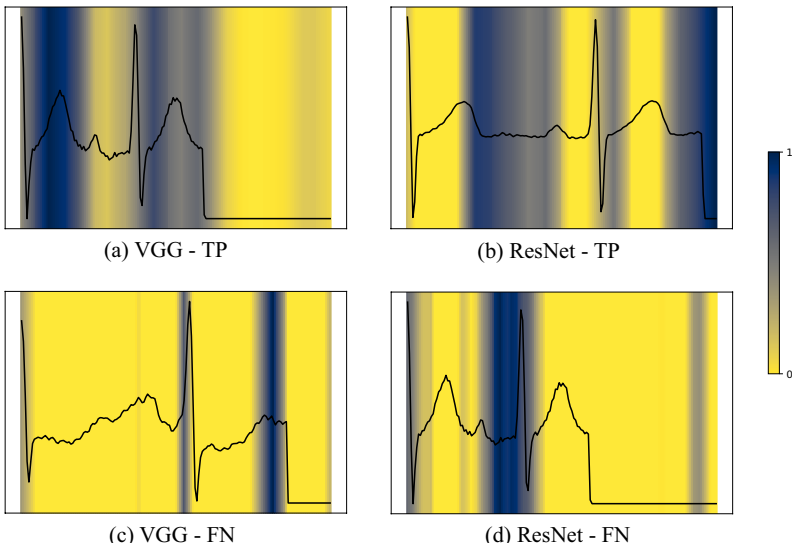


Fig. 11 Grad-CAM visualization with dark shadow for *TP* and *FN* cases classified by VGG and ResNet models

physicians to accurately classify heartbeats categories from ECGs. By applying techniques that provide clear and understandable insights into the decision-making process of DL algorithms, these methods bridge the gap between complex DL models and the expertise of physicians. Through visualizations and feature importance explanations, physicians can recognize the specific patterns and markers that lead to arrhythmia detection. This transparency not only enhances trust in AI systems but also enables medical professionals to validate and fine-tune the algorithms, making them more reliable and efficient. As a result, explainable AI empowers physicians with a comprehensive understanding of how such models arrive at their conclusions, facilitating more accurate and confident diagnoses of heart arrhythmias.

4 Conclusion, limitations, and future work

In summary, this paper validates the effectiveness of two distinct DL models in classifying five ECG heartbeat categories. By extracting heartbeats from ECG recordings and employing them as inputs for classification algorithms, we achieved two significant contributions. Firstly, we demonstrated the applicability of DL models using a transfer learning approach on entirely unseen test sets. Our experimental results showcase the models' ability to generalize and perform well on new datasets, with AUC-ROCs exceeding 0.961. Secondly, we delved into the classification outcomes of these complex DL models, employing the GRAD-CAM technique for explanation. This step ensures comprehensibility for non-technical users, a crucial step towards implementing such diagnostic tools in clinical practice. The findings of this research open avenues for physicians to utilize these models *in house*, enhancing the diagnosis and monitoring of heart patients. Moving forward, future research can explore the extension of these methodologies to broader healthcare applications, as well as refining the interpretability of DL models for enhanced clinical decision-making.

While the study presents a promising approach for automated analysis of ECGs using deep transfer learning, several limitations should be acknowledged.

1. Dataset limitations: the evaluation of the deep neural networks relies on two specific ECG datasets (SVDB and INCARTDB). The generalizability of the models may be limited by the characteristics and variability of these datasets, and their performance may vary when applied to other datasets with different demographics or recording conditions.
2. *Model complexity*: while the use of deep transfer learning is advantageous for leveraging pre-trained models and achieving high classification performance, it also introduces complexity. The DL architectures (VGG and ResNet) have a large number of parameters, increasing computational demands.
3. *Clinical validation*: although the models achieve high performances on the test datasets, their performance in real-world clinical settings remains to be validated. Factors such as data variability and the presence of confounding factors may affect model performance in practice.
4. *Practical implementation*: the feasibility of deploying these models for real-time monitoring of heart patients using ECG data needs to be addressed. Considerations such as computational resources, integration with existing clinical workflows, and regulatory compliance are crucial for successful implementation in clinical settings.

In summary, while the proposed approach shows promise for automated ECG analysis, further research, such as a longitudinal study, is needed to address the above-mentioned limitations and ensure the robustness, generalizability, and practical utility of the models in real-world clinical practice. Future research can explore the extension of these methodologies to broader healthcare applications, as well as refining the interpretability of DL models for enhanced clinical decision-making. Specifically, investigating the adaptation of these models to diverse demographic groups and varying recording conditions would strengthen their applicability across different patient populations. Moreover, enhancing the robustness of DL architectures by optimizing their computational efficiency without compromising performance is crucial for practical implementation in clinical settings. Additionally, longitudinal studies are essential to validate the sustained performance and reliability of these models over extended periods, addressing concerns about their efficacy in real-world scenarios. Finally, collaboration with clinicians and healthcare providers is paramount to ensure seamless integration into existing workflows and adherence to regulatory standards, facilitating the eventual adoption of these technologies as valuable tools in routine clinical practice.

Funding Open access funding provided by University of Southern Denmark.

Data availability The datasets used in this work are publicly available on PhysioNet data repository, https://www.physionet.org/content/mitdb/1.0.0/, https://www.physionet.org/content/incartdb/1.0.0/, and https://www.physionet.org/content/svdb/1.0.0/.

References

Ahsan MM, Uddin MR, Ali MS, Islam MK, Farjana M, Sakib AN, Al Momin K, Luna SA (2023) Deep transfer learning approaches for monkeypox disease diagnosis. Expert Syst Appl 216:119483

Alonso-Atienza F, Rojo-Álvarez JL, Rosado-Muñoz A, Vinagre JJ, García-Alberola A, Camps-Valls G (2012) Feature selection using support vector machines and bootstrap methods for ventricular fibrillation detection. Expert Syst Appl 39(2):1956–1967

Andersen RS, Peimankar A, Puthusserypady S (2019) A deep learning approach for real-time detection of atrial fibrillation. Expert Syst Appl 115:465–473

Baeza-Yates R, Ribeiro-Neto B (1999) Modern information retrieval, vol 463. ACM Press, New York

Chen S-W, Clarkson PM, Fan Q (1996) A robust sequential detection algorithm for cardiac arrhythmia classification. IEEE Trans Biomed Eng 43(11):1120–1124

Chen H, Fang X, Fang H (2022) Multi-task prediction method of business process based on Bert and transfer learning. Knowl-Based Syst 254:109603

Cheplygina V, Pena IP, Pedersen JH, Lynch DA, Sørensen L, De Bruijne M (2017) Transfer learning for multicenter classification of chronic obstructive pulmonary disease. IEEE J Biomed Health Inform 22(5):1486–1496

Chollet F et al (2015) Keras. https://keras.io

Das MK, Ari S (2014) ECG beats classification using mixture of features. Int Schol Res Not 2014(1):178436

De Chazal P, Reilly RB (2006) A patient-adapting heartbeat classifier using ECG morphology and heartbeat interval features. IEEE Trans Biomed Eng 53(12):2535–2543

Engin M (2004) Ecg beat classification using neuro-fuzzy network. Pattern Recogn Lett 25(15):1715–1722

Faezipour M, Saeed A, Bulusu SC, Nourani M, Minn H, Tamil L (2010) A patient-adaptive profiling scheme for ECG beat classification. IEEE Trans Inf Technol Biomed 14(5):1153–1165

Fan C, Sun Y, Xiao F, Ma J, Lee D, Wang J, Tseng YC (2020) Statistical investigations of transfer learning-based methodology for short-term building energy predictions. Appl Energy 262:114499

Gaur P, Malaviya V, Gupta A, Bhatia G, Pachori RB, Sharma D (2022) Covid-19 disease identification from chest ct images using empirical wavelet transformation and transfer learning. Biomed Signal Process Control 71:103076

Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE (2000) Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. Circulation 101(23):215–220

González-Vidal A, Mendoza-Bernal J, Niu S, Skarmeta AF, Song H (2022) A transfer learning framework for predictive energy-related scenarios in smart buildings. IEEE Trans Ind Appl 59(1):26–37

Greenwald SD, Patil RS, Mark RG (1990) Improved detection and classification of arrhythmias in noise-corrupted electrocardiograms using contextual information. IEEE

Güler I, Übeylı ED (2005) ECG beat classifier designed by combined neural network model. Pattern Recogn 38(2):199–208

Guo L, Sim G, Matuszewski B (2019) Inter-patient ECG classification with convolutional and recurrent neural networks. Biocybern Biomed Eng 39(3):868–879

He H, Bai Y, Garcia EA, Li, S (2008) Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence) . IEEE, pp 1322–1328

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778

Himeur Y, Elnour M, Fadli F, Meskin N, Petri I, Rezgui Y, Bensaali F, Amira A (2022) Next-generation energy systems for sustainable smart cities: Roles of transfer learning. Sustainable Cities and Society 85:104059

Hua Y, Sevegnani M, Yi D, Birnie A, McAslan S (2022) Fine-grained rnn with transfer learning for energy consumption estimation on evs. IEEE Trans Industr Inf 18(11):8182–8190

Ismail Fawaz H, Forestier G, Weber J, Idoumghar L, Muller P-A (2019) Deep learning for time series classification: a review. Data Min Knowl Disc 33(4):917–963

Jahan MS, Mansourvar M, Puthusserypady S, Wiil UK, Peimankar A (2022) Short-term atrial fibrillation detection

using electrocardiograms: a comparison of machine learning approaches. Int J Med Informat 163:104790
Kallas M, Francis C, Kanaan L, Merheb D, Honeine P, Amoud H (2012) Multi-class svm classification combined with kernel pca feature extraction of ECG signals. In: 2012 19th international conference on telecommunications (ICT). IEEE, pp 1–5
Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980
Kumar S, Mallik A, Kumar A, Del Ser J, Yang G (2023) Fuzz-clustnet: coupled fuzzy clustering and deep neural networks for arrhythmia detection from ECG signals. Comput Biol Med 153:106511
Lal A, Kumar P, Halder S (2023) Heartbeat classification based on deep convolutional neural network. In: 2023 international conference on networking and communications (ICNWC). IEEE, pp 1–4
Lassoued H, Ketata R (2018) ECG multi-class classification using neural network as machine learning model. In: 2018 international conference on advanced systems and electric technologies (IC_ASET). IEEE, pp 473–478
Li X, Xie H, Lau RY, Wong T-L, Wang F-L (2018) Stock prediction via sentimental transfer learning. IEEE Access 6:73110–73118
Liu J, Zhang Q, Li X, Li G, Liu Z, Xie Y, Li K, Liu B (2021) Transfer learning-based strategies for fault diagnosis in building energy systems. Energy Build 250:111256
Luz EJdS, Schwartz WR, Cámara-Chávez G, Menotti D (2016) ECG-based heartbeat classification for arrhythmia detection: a survey. Comput Methods Program Biomed 127:144–164
Melgani F, Bazi Y (2008) Classification of electrocardiogram signals with support vector machines and particle swarm optimization. IEEE Trans Inf Technol Biomed 12(5):667–677
Mohammad U, Saeed F (2022) Spertl: epileptic seizure prediction using eeg with resnets and transfer learning. In: 2022 IEEE-EMBS international conference on biomedical and health informatics (BHI). IEEE, pp 1–5
Mondéjar-Guerra V, Novo J, Rouco J, Penedo MG, Ortega M (2019) Heartbeat classification fusing temporal and morphological information of ECGs via ensemble of classifiers. Biomed Signal Process Control 47:41–48
Moody GB, Mark RG (2001) The impact of the mit-bih arrhythmia database. IEEE Eng Med Biol Mag 20(3):45–50
Mousavi S, Afghah F (2019) Inter-and intra-patient ECG heartbeat classification for arrhythmia detection: a sequence to sequence deep learning approach. In: ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 1308–1312
Murat F, Yildirim O, Talo M, Baloglu UB, Demir Y, Acharya UR (2020) Application of deep learning techniques for heartbeats detection using ECG signals-analysis and review. Comput Biol Med 120:103726
Peimankar A, Puthusserypady S (2021) DENS-ECG: a deep learning approach for ECG signal delineation. Expert Syst Appl 165:113911
Peimankar A, Jajroodi MJ, Puthusserypady S (2019) Automatic detection of cardiac arrhythmias using ensemble learning. In: TENCON 2019-2019 IEEE region 10 conference (TENCON). IEEE, pp 383–388
Peimankar A, Puthusserypady S (2018) Ensemble learning for detection of short episodes of atrial fibrillation. In: 2018 26th European signal processing conference (EUSIPCO). IEEE, pp 66–70
Peimankar A, Puthusserypady S (2019) An ensemble of deep recurrent neural networks for p-wave detection in electrocardiogram. In: ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 1284–1288
Sannino G, De Pietro G (2018) A deep learning approach for ECG-based heartbeat classification for arrhythmia detection. Futur Gener Comput Syst 86:446–455
Sellami A, Hwang H (2019) A robust deep convolutional neural network with batch-weighted loss for heartbeat classification. Expert Syst Appl 122:75–84
Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision, pp 618–626
Sharma P, Dinkar SK, Gupta D (2021) A novel hybrid deep learning method with cuckoo search algorithm for classification of arrhythmia disease using ECG signals. Neural Comput Appl 33:13123–13143
Shi H, Wang H, Zhang F, Huang Y, Zhao L, Liu C (2019) Inter-patient heartbeat classification based on region feature extraction and ensemble classifier. Biomed Signal Process Control 51:97–105
Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556
Testing and reporting performance results of cardiac rhythm and ST segment measurement algorithms. https://webstore.ansi.org/standards/aami/ansiaamiec572012r2020. Accessed: 2023-07-07
Tihonenko V, Khaustov A, Ivanov S, Rivin A, Yakushenko E (2008) St petersburg incart 12-lead arrhythmia database. PhysioBank PhysioToolkit and PhysioNet
World Health Organization (WHO). https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). Accessed 07 July 2023
Wu D, Wang X, Wu S (2022) Jointly modeling transfer learning of industrial chain information and deep learning for stock prediction. Expert Syst Appl 191:116257
Ye C, Kumar BV, Coimbra MT (2012) Heartbeat classification using morphological and dynamic features of ECG signals. IEEE Trans Biomed Eng 59(10):2930–2941
Yıldırım Ö, Pławiak P, Tan R-S, Acharya UR (2018) Arrhythmia detection using deep convolutional neural network with long duration ECG signals. Comput Biol Med 102:411–420
Yu J, Qiu M, Jiang J, Huang J, Song S, Chu W, Chen H (2018) Modelling domain relationships for transfer learning on retrieval-based question answering systems in e-commerce. In: Proceedings of the 11th ACM international conference on web search and data mining, pp 682–690
Zhu H, Samtani S, Chen H, Nunamaker JF Jr (2020) Human identification for activities of daily living: a deep transfer learning approach. J Manag Inf Syst 37(2):457–483