Bıçakcı Yeşilkaya and Guest
EURASIP Journal on Information Security (2025) 2025:7
https://doi.org/10.1186/s13635-025-00193-8

EURASIP Journal on Information Security

RESEARCH **Open Access**

Activity-aware electrocardiogram biometric verification utilising deep learning on wearable devices

Hazal Su Bıçakcı Yeşilkaya[1*] and Richard Guest[2]

Abstract

With the advancement of technology and the increasing use of wearable devices, information security have become a necessity. Although many biometrics authentication methods have been studied on these devices to ensure information security, an activity-aware deep learning (DL) model that is compatible with different device types and uses only electrocardiogram signals has not been studied. Our objective is to investigate DL models that exclusively use ECG signals during several physical activities, facilitating their implementation on various devices. Through this research, we aim to contribute to the advancement of wearable devices for the purpose of biometric verification. In this context, this study investigates the application of adaptive techniques that rely on prior activity classification to potentially improve biometric performance using DL models. In this study, we compare three time-frequency representations to generate images for activity classification using GoogleNet, ResNet50 and DenseNet201, and for biometric verification using ResNet50 and DenseNet201. Despite employing various convolutional neural network (CNN) models, we could not achieve high accuracy in activity classification. Consequently, manually classified samples were used for activity-aware biometric verification. We also provide a detailed comparison of various DL parameters. We use a public dataset simultaneously collected from both medical and wearable devices to offer a cross-device comparison. The results demonstrate that our method can be applied to both wearable and medical devices for activity classification and biometric verification. Besides, although it is known that DL requires a large amount of training data, our model, which was created using a small amount of training data and a real-life biometric verification scenario, achieved comparable results to studies using a large amount of data. The model was achieved 0.16% to 30.48% better results when classified according to their physical activities.

Keywords ECG biometrics, Activity classification, Biometric authentication, Wearable devices

1 Introduction

As technological advancements improve the growth and necessity of online transactions, the demand for biometric authentication models to protect against potential security breaches during these operations has concurrently increased. To increase security, biometrics traits are used in addition to systems such as classical PINs and passwords. An electrocardiogram (ECG), a technique that records the heart's electrical activity via electrodes attached to the human body, can discern the distinct cardiac rhythm inherent to each individual. This unique characteristic positions ECG as a candidate for biometric authentication systems [1, 2].

The biometric terms mentioned in this study are explained by ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) in ISO/IEC 2382-37:2022(E)

*Correspondence:
Hazal Su Bıçakcı Yeşilkaya
hazalsu.bicakci@bakircay.edu.tr; hazalsubicakci@gmail.com
[1] Department of Electrical and Electronics Engineering,1Zmir BakıRçAy University, Seyrek Campus, Menemen 35665, İzmir, Türkiye
[2] School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, Hampshire, UK

[3]. According to ISO/IEC 2382-37:2022(E) [3], authentication is the process of proving the authenticity of an entity, though it should not be used interchangeably with biometric verification or identification, as "biometric recognition" is the preferred term. Biometric identification involves searching a database to match biometric data with a specific individual, while biometric verification confirms a claimed identity by comparing the provided biometric data with stored reference data.

According to 2018 data in the UK, 900,000 people were living with heart failure [4], while heart-related complaints were the second most common emergency calls in the USA in 2019 [5]. Cardiac health is examined with medical 12-lead ECG device measurement and this method is 70 years old [6]. Since medical devices are not designed for use in daily life, wearable technologies have begun to be used frequently. Many mobile health tracking applications have started to be used [7], along with the frequent use of wearable devices such as smartwatches [7] and chest bands [8]. This raised the issue of whether wearable technology provides as reliable results as medical devices [5]. In addition, given that ECG signals can be influenced by various physical activities and emotional states [9], it is essential to investigate the performance of wearable devices under different conditions. Therefore, the performance of medical and wearable devices is a frequently researched topic due to their different configurations and intended uses [5, 9–11].

In order to create ECG-based applications and compare devices, the characteristics of the ECG waveform must first be examined. Each heartbeat of a healthy person has P, R and T peaks and Q and S troughs, and an example heartbeat is shown in Fig. 1.

While the distance between these peaks and troughs is generally consistent for healthy individuals, distorted peaks and unbalanced time intervals indicate an irregular and unhealthy heart rhythm [12]. The figure illustrates the three distinct phases of each cardiac cycle. The initial phase involves the depolarisation of the atria, which is represented by the P wave. The subsequent phase is ventricular depolarisation, which occurs by signals transmitted to the Hiss-Purkinje systems and is denoted by the Q, R and S waves. The final phase contains ventricular repolarisation and is signified by the T wave. The cardiac cycle is constantly repeated, but its stability changes due to reasons such as physical activities, drug use or heart disease [12].

The performance of ECG-based biometric systems is adversely affected by environmental, biological or physiological factors that affect ECG signals. These effects must be reduced in order for these systems to be consistent under all conditions and to obtain reliable results. To reduce these effects, a biometric verification framework is proposed that recommends activity classification before the biometric verification model and operating biometric verification models for each activity class. This study aims to explore deep learning (DL) models that exclusively utilise ECG signals across various physical

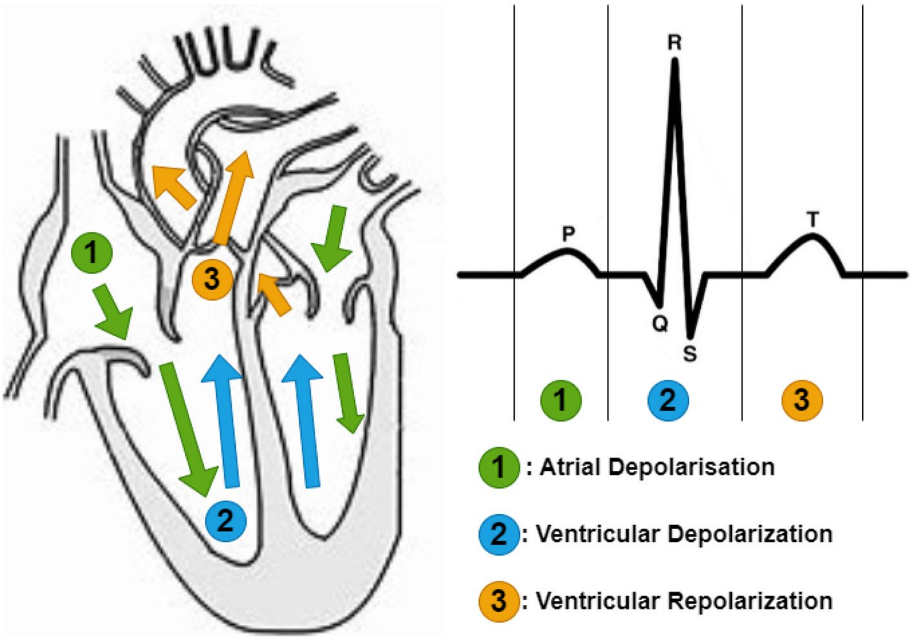

Fig. 1 A single cardiac cycle and ECG signal with the representation of P, Q, R, S and T peaks

activities, enabling their integration into diverse devices. Furthermore, it seeks to evaluate the activity-aware biometric verification framework, previously assessed using classical machine learning (ML) models [9], in comparison with DL models, thereby contributing to the advancement of wearable technologies in biometric verification.

In this framework, we classified activities prior to biometric verification. Each activity class is assessed in the biometric verification task separately. ResNet50, DenseNet201 and GoogleNet 2D convolutional neural networks (CNN) models were used for activity classification and biometric verification. In these CNN models, spectrogram, Mel-spectrogram and scalogram images were tested using different window sizes. We experimented with different CNN models for activity classification but faced challenges in achieving high accuracy. Therefore, we resorted to using manually classified samples for activity-aware biometric verification. A medically approved device and a consumer-based device were compared using simultaneously collected ECG data. In addition to comparing detailed DL parameters and the verification framework, this study's ability to obtain successful results in short authentication times by taking into account the real-life scenario in biometric verification models is our contribution to science.

A brief examination of the contributions of DL models in this study reveals the following: both devices outperformed ML models [9] in biometric verification tasks within DL frameworks but demonstrated limited effectiveness in activity classification. In activity classification tasks, the Faros device achieved results comparable to classifiers such as Decision Trees (DT), support vector machines (SVM), and k-nearest neighbors (KNN) in ML studies [9]. However, the Hexoskin device achieved higher accuracy rates in activity classification within ML models [9].

2 Related works

Related works of this study include the utilisation of ECG data for biometric verification and identification, the classification of physical activities and the exploration of activity-aware biometric systems via DL models.

2.1 ECG biometric verification and identification

It has been observed that devices, enrollment times, datasets and applied DL parameters achieve different performances in biometric authentication models and have been frequently researched in the literature. For instance, Li et al. [13] compared biometric identification performances with the cascaded CNN model using different datasets. The cascaded CNN model was created by using two CNN models for training: F-CNN (a 1 CNN model is called F-CNN because it is created for feature extraction) and M-CNN (a 1 CNN model is called M-CNN because it is created for matching in biometric identification). Unlike our study, healthy and unhealthy subjects were used together in their study. R-peaks segments were used to feed the CNNs. They expressed that F-CNN is useful for multi-class classification. However, F-CNN can be easily affected by data variance. Even if they compared 5 public datasets in their experiments, they utilised the same dataset separately to train models because using multiple or merged datasets could increase the variance. M-CNN was used with raw ECG signals. When they fed the M-CNN with R-peak segments, the results were worse than raw signals. F-CNN was used to learn features and these features were used to feed M-CNN. The features of one current heartbeat and one template from F-CNN were combined to crerate input for M-CNN. M-CNN was performed for binary matching. In this way, they achieved higher identification accuracies (from 89.1 to 93.1%). Although the datasets and CNN structures they used were different from ours, they tested data from 1 to 20 heartbeats and found that the identification accuracy rate improved as the number of heartbeats used in the test increased. However, they determined that using 3 heartbeats in the test was optimal when considering the time cost. This study enabled us to identify the optimal number of samples required for biometric verification.

AlDuwaile and Islam [14] also used R-peaks which were calculated from a 0.5-s window segment in their CNN biometric recognition system. P-peaks and R-peaks segments were used with continuous wavelet transformation (CWT) to create images. They compared GoogleNet, ResNet, EfficientNet and MobileNet with different ECG time windows which were selected from blindly and peak segmented images. One hundred subjects from PTB and 90 subjects from ECG-ID datasets were used in their study. In blind segmentation, the 2-s time window was selected as the best-performed window size with 98.14% identification accuracy rates using GoogleNet. In heartbeat segmentation, 0.5-s time window (single heartbeat) heartbeat segmented data have higher identification accuracy than other cases. The lowest half total of EER (i.e. (false reject rate (FAR)+false accept rate (FAR))/2) was achieved by ResNet and GoogleNet however, they did not mention any imposter samples or real EERs. This study showed the effect of different window sizes on a few CNN models. Although it used different datasets than our study for biometric identification purposes, it influenced us to investigate different window segments and different CNN models.

Begum et al. [15] compared 4 distinct DenseNet CNN architectures with several training and testing sample sizes on 8 ECG datasets for biometric identification purposes. These datasets include healthy and unhealthy subjects with a range of subject sizes. The structure of 4 different CNN architectures were as follows: Architecture #1 had 5 convolutional layers, 3 concatenation layers, and 3 filters in each convolutional layer. Architecture #2, with the same number of convolutional and concatenation layers but with 16 and 5 filters in each convolutional layer. Architecture #3, which had 6 convolutional layers, 4 concatenation layers, and 10 filters in each convolutional layer. Lastly, Architecture #4, with 6 convolutional layers, 5 concatenation layers, and 10 filters in each convolutional layer. The test prediction accuracies for 4 different architectures were expressed in Table 1. This study influenced our research by highlighting the significance of CNN parameters and identifying the optimal training and testing ratio.

Abd El-Rahiem and Hammad [39] used combined features which were extracted from spectrogram images using VGG-16, VGG-19, AlexNet, ResNet50, ResNet101 and GoogleNet CNNs. However, they used SVM and KNN classifiers for biometric authentication stages. In the MWM-HIT dataset, 10 s of data from each of the activities of sitting, standing, supine, exercise sitting and exercise standing were used. However, no research has been conducted on the effects of the activities. In addition, unbalanced genuine and imposter samples were used in authentication tasks. Using unbalanced genuine and imposter samples in biometric authentication can cause the model to become biased, leading to inaccurate results. The model might overfit to the majority class, and it can distort performance metrics and increase security risks by failing to correctly identify imposters. Byeon et al. [16] used each R-peak's scalogram images to feed AlexNet, GoogleNet and ResNet CNN structures. Although they compared different optimisers, CNN models, mini-batch sizes and transfer learning parameters, the effect of the parameters on the performance could not be observed because they used different parameters in the biometric verification models. In addition, although the results are prominent, they do not reflect a realistic verification scenario because they were made using only 1 genuine subject and 60 imposter subjects (i.e. single-user authentication). In biometric verification, testing only one genuine subject is unrealistic or widely applicable. Instead, testing multiple subjects separately and averaging the results for each subject makes the verification scenario more generally applicable. Byeon and Kwak [40] used the single heartbeat spectrogram, log-spectrogram, Mel-spectrogram, scalogram and Mel frequency cepstrum coefficient (MFCC) to feed VGGNet19, ResNet101, DenseNet201 and Xception CNN models and compared their biometric identification performances across different datasets. For instance, in the PTB-ECG dataset, the ResNet101 model outperformed the DenseNet201 model regarding test accuracy. However, the opposite was observed in the other dataset. Furthermore, the effectiveness of the image representations also varied across datasets. They found the best identification rates when they used Xception CNN and MFCC images. In the PTB-ECG dataset, the spectrogram was the most successful, followed by the scalogram and Mel-spectrogram. In contrast, in the other dataset, the scalogram was the most successful, followed by the spectrogram and Mel-spectrogram. Although the study [40] supports our study in these points, it differs from our studies in that it does not include activities and does not perform biometric verification.

Ciocoiu and Cleju [19] used UofT and CYBHi datasets, they used S-transform plots of a single beat, Gramian Angular Fields, Phase-Space Trajectories, and Recurrence plots to feed 10 different CNN models. ResNet50 achieved the highest identification rates. In addition, they used their own 2-D CNN model for verification. They trained the model with 52 subjects (700 segments per subject) during the sitting activity from the UofT dataset and tested with 200 subjects (200 segments per subject) from the UofT dataset. During the training phase for the identification task, all image types were used to train the system. However, during verification tasks, only the S-Transform was utilised as it has shown better results compared to other methods. While there was no activity classification in their study, they referred to it as an open challenge.

In addition to these studies, other state-of-the-art studies in the literature are also shown in Table 1. These studies applied different methods for biometric verification and identification cases. Although a one-to-one comparison between the studies cannot be made because of the different methods they used, they inspired our work because they investigated different CNN models, DL parameters, different image types and template sizes. However, although these studies inspired us, the dataset we used is not available in any research. In addition, even if they use data collected during different physical activities, the biometric verification framework we created has not been studied in any of them. No previous study has compared DL parameters in biometric verification performance after activity classification using the the Vollmer dataset (Simultaneous physiological measurements with five devices at differ- ent cognitive and physical loads [41]).

2.2 Activity classification and activity-aware biometric models

Some studies have investigated activity recognition using mobile and wearable devices before biometric user

Table 1 State-of-the-art studies in biometric authentication using DL

Authors	Datasets	# of Subjects	Recordings	Methods	Performances
Byeon et al. (2019) [16]	PTB [17] CU [18]	290 100	50% training (0.9 or 0.75 training and 0.1 or 0.25 validation), 50% testing data	R-peak segmentations, scalogram images, 1 subject as a genuine, 60 subjects as imposters	ResNet: 0% EER, 0% FAR and FRR, 100% accuracy rate GoogleNet: 0% EER and FRR, 2% FAR, 99.17% accuracy rate
Ciocoiu et al. (2020) [19]	UofTDB [20] CYBHi [21]	1020 65	700 segments in training (52 subjects) for UofT, 200 segments in training for CYBHi, 1, 3, 5 and 13 templates for testing	S-transform plots of a single beat, Gramian Angular Fields, Phase-Space Trajectories, and Recurrence plots to feed 2-D CNN, AlexNet, SqueezeNet, GoogleNet, VGG16, MobileNetv2, Inceptionv3, ResNet50, DenseNet201 and Xception, Euclidean distance matching	EERs: single template: 9.69%, 3 templates: 5.48%, 5 templates: 4.86% and 13 templates: 4.4%
Li et al. (2020) [13]	FANTASIA [22] CEBSDB [23] NSRDB [24] STDB [25] AFDB [26]	40 20 18 28 23	196 heartbeats per subject (total: 2400 samples training, 8000 validation and 8000 testing samples)	250 Hz resampled sampling frequency Fixed R peak locations (79th samples) F-CNN, M-CNN and Cascade CNN	Identification accuracies: F-CNN: 97.8% M-CNN: 98.1% Cascade CNN: 99.3%
Alduwaile et al. (2021) [14]	PTB [17] ECG-ID [27]	100 90	PTB: 10,000 heartbeat segments ECG-ID: 9000 heartbeat segments	0.5, 0.75, 1, 1.5, 2, 2.5 and 3 s signal segments. Blind, R-centered, P-P, R-R segmentations, GoogleNet, ResNet, EfficientNet, MobileNet and Small CNNs	Identification accuracies: ResNet and MobileNet : 100%, GoogleNet: 99.90% EERs: ResNet: 2.48%, Small CNN: 3.33%, GoogleNet: 3.7%, EfficientNet: 5.17%, MobileNet: 5.84%
Hammad et al. (2021) [28]	PTB [17] CYBHi [21]	290 63	70% training, 20% validation and 10% testing data	2-s time windowed samples, 10-fold cross-validation method for authentication, 5 distinct 1-D ResNet CNN architectures, PTB: 300 genuine/199 imposter samples CYBHi: 240 genuine/120 imposter samples	EERs: 1-D ResNet: PTB: 1.53% and CYBHi: 0.27% ResNet-attention: PTB: 1.39% and CYBHi: 0.68%
Begum et al. (2022) [15]	MIT-BIH Arrhythmia [29] NSRDB [24] PTB [17] QTDB [30] ECG-ID [27] IAF [31] CU [18] MIMIC-II/III [32, 33]	10 13 38 13 25 7 20 124	Varied training/testing ratios: from 90%/10% to 10%/90%	500-Hz resampled sampling frequency 1000 samples per template/image Created 4 DenseNet architectures used in identification	The best results: 90% training/10% testing, 250 subjects has higher accuracy than 120 subjects identification accuracies: architecture 1: 99.42% architecture 2: 99.84% architecture 3: 99.80% architecture 4: 99.94%
Rahiem et al. (2022) [39]	MWM-HIT [34] PTB [17]	100 290	10-s recordings × 5 trials, 1–7 signals per patient in PTB	10-fold cross-validation method for authentication, spectrogram images VGG16, VGG19, AlexNet, ResNet50, ResNet101 and GoogleNet CNNs to create features, Unbalanced genuine and imposter samples	Authentication accuracy rates: MWM-HIT: SVM 99.4%, KNN: 99% PTB Internal fusion: 99.6%

Table 1 (continued)

Authors	Datasets	# of Subjects	Recordings	Methods	Performances
Melzi et al. (2023) [35]	In-house [36, 37]	81974	5162 genuine/25801 imposter	Single-lead CNN and Euclidean distance matching	In-house: 5.55% EER
	PTB [17]	165	113 genuine/565 imposter		PTB: 5.12% EER
	ECG-ID [27]	90	89 genuine/445 imposter		ECG-ID: 0.26% EER
	CYBHi [21]	65	63 genuine/315 imposter		CYBHi: 5.44% EER
Prakash et al. (2023) [38]	ECG-ID [27]	90	20 beats per person	Siamese Network	Accuracy rates: single beat image: 91% dual beat image: 99.85% triple beat image: 99.90%

identification and verification [42–45]. These studies are shown in Table 2. However, in addition to ECG signals, most also use gyroscope or accelerometer data for activity classification. Although many studies have been conducted using these sensors that indicate movement and speed in activity classification, classifying activity using only ECG signals is one of the challenging tasks. In addition, using less sensor data reduces computational cost and processing time. For instance, Liu et al. [45] used ECG signals and accelerometer data that were collected from a wearable chest sensor. FIR low-pass and high-pass filters were used to eliminate noisy components. Time-domain features, DC features and energy features were used to classify standing, sitting, lying, sitting, walking and coughing. Activity recognition results are a 2.4% detection error rate (DER) for standing, 0.0% DER for lying, 4.9% DER for sitting, 2.4% DER for coughing, 5.6% DER for sitting down and 3.2% DER for squatting down. The main difference between their study and ours is that they use accelerometer data to classify activities.

Butt et al. [46] collected ECG signals from the wearable device to classify falling, daily activities and resting. They collected data from 8 subjects and compared initial learning rates, stochastic gradient descent with momentum (SGDM) and RMSProp optimisers. They split data 80%−10%−10% and 60%−20−20% for training, testing and validation, respectively. They stated that classification accuracy rates generally increase as the initial learning rate becomes smaller, the RMSProp optimiser gives better results than SGDM and in the case of 80% training has higher accuracy rates than the 60% training case. In addition, AlexNet had better results than GoogleNet. This study inspired us because it aimed to classify activities using only ECG signals and employed scalogram images in DL models. However, the activities they examined and the methods they utilised differ from our study. Additionally, our research investigates spectrogram and Mel-spectrogram images.

Cosoli et al. [47] used the chest-worn (Zephyr Bio-Harness 3.0) device and smartwatch (Samsung Galaxy Watch3) performances for ECG-based activity classification. They classified resting, walking, slow running and running activities with several ML and DL methods. Although this study tested different ML models with HR-based features, it only performs activity classification and does not perform biometric verification as in our study.

Kim et al. [44] used finger and limb electrodes to collect ECG data from 104 subjects. They compared biometric verification performances during sitting, standing and exercise activities. They used stationary wavelet transform (SWT) and infinite feature selection (Inf-FS) methods to create a feature vector. The SWT, discrete wavelet transform (DWT), short-time Fourier transform (STFT) and AC/LDA methods were used for biometric verification. In addition, they addressed that sitting and standing activities have lower EERs than the exercise activity. Unlike our study, this research did not use activity classification. Instead, it used 125 heartbeats for enrollment and 125 heartbeats for testing in biometric verification. Since biometric verification needs to be quick in real-life scenarios, the time required for verification in the 125-beat recording and test case is very long and unrealistic.

For the biometric authentication task, Nawawi et al. [48] used the Hexoskin wearable device to collect data during walking, standing and sitting. They extracted QRS-segmented fiducial features and used a quadratic-SVM classifier. They compared different training and testing data sizes and reported that the optimum values were 80% training and 20% testing data. Although the numbers and EER rates of genuine and imposter samples are not specified, FAR and FRR rates are stated. If we consider the FAR and FRR rates when the training and test data are selected from the same activity, 20% FRR and 0.51% FAR in the standing activity, 12.22% FRR and 1.37% FAR in the sitting activity, and 3.64% FRR and 0.93% FAR in the walking activity are observed. The fact that FAR and FRR ratios are quite different from each other shows that the numbers of genuine and imposter samples used are not equal and the reliability of the system is low.

Wahabi et al. [43] worked on posture classification before biometric verification. These postures are sitting, standing, resting and tripod/squat position. Data from 52 subjects from the UofTDB dataset were used in their study. Biometric verification was performed using DWT feature extraction, AC/LDA method and SVM classifier. A mean of 1.50% EER was obtained when the same postures were used for testing and enrollment, while 8.24% EER was obtained in different posture situations. It differs from our study because it does not have the same amount of records for each activity, it only performs posture classification and it is not known with which labels the data is passed to biometric verification after posture classification (i.e. with only original posture labels or including wrongly classified posture labels). Moreover, in another study, Wahabi et al. [42] added an exercise activity to the existing postures. In the study [42], DWT, Time-frequency content method, EigenPulse method and AC/LDA method were compared. It was stated that an average of 69 of 1020 subjects were used for each activity. The number of individuals involved in various activities during data collection varies. For instance, the number of subjects varied from 63 in supine and tripod postures to 1020 in the sitting position. For this reason, it was observed that there were not equal numbers of subjects or samples for each activity. This study stated

Table 2 State-of-the-art studies in activity-aware biometric models

Authors	Datasets	# of subjects	Recordings	Methods	Performances	
					Accuracy rates in activity classification	EERs
Wahabi et al. (2014) [42]	UofTDB [20]	1020	69 subjects 50% training/50% testing 5 activities: sitting, standing, exercise, supine and tripod	DWT features, time-frequency content, EigenPulse and AC/LDA methods	52–96%	DWT: 2.62–32.87% time-frequency content: 1.58–33.82% EigenPulse: 11.06–39.74% AC/LDA: 1.44–24.10%
Wahabi et al. (2015) [43]	UofTDB [20]	1020	52 subjects 50% training/50% testing 5 activities: sitting, standing, exercise, supine and tripod	DWT features, AC/LDA and SVM classifiers	Tripod: 94.12%, sitting: 98.04%, stranding: 98.04%, resting: 98.04%	Same posture: 1.50% different postures: 8.24%
Liu et al. (2018) [45]	Private	13	Accelerometer and ECG total: 134.1 min 5 activities: Standing, sitting, squatting, coughing and walking	64 sample windows with 32 samples overlapping (each window is 1.28 s), frequency and time domain features		Decision Trees (DTs): mean 3.08% DER
Kim et al. (2019) [44]	Private	104	70 s × 5 trials (at least 250 beats per person) 3 activities: sitting,standing and exercise	SWT and Inf-FS features, 125 beats for enrollment and 125 beats for testing, DWT, STFT and AC/LDA methods		SWT: 0–5.61% DWT: 1.77–35.38% STFT: 0.42–36.55% AC/LDA: 0.94–25.74%
Butt et al. (2021) [46]	Private	8	Varied recording time 80% training/ 20% testing and 60% training/40% testing 4 activities: laying, rolling, falling down and daily activities	Scalogram images, AlexNet and GoogleNet CNN	AlexNet: 99.2% GoogleNet: 98.4%	
Cosoli et al. (2023) [47]	Private	30	30 s × 4 activities 70% training/30% testing 4 activities: resting, walking, slow running and running	ML with HR-based features and DL classifiers	RF: 81% LSTM: 79% SVM: 78% NB: 76% simple logistic: 70% DTs: 66%	
Nawawi et al. (2023) [48]	Private	11	15 min (80% training/20% testing) 3 activities: standing, sitting and walking	QRS segments as features, SVM classifier 1 subject as genuine and 10 subjects as imposters		Standing: 20% FRR, 0.51% FAR sitting: 12.22% FRR, 1.37% FAR walking: 3.64% FRR, 0.93% FAR

that static activities (such as standing, supine and sitting) obtained lower EER and higher activity classification accuracy rates results than dynamic activities (tripod and exercise). The studies [42, 43], form the basis of our proposed framework, using fingertip ECG data (mobile sensor) and comparing different methods. However, it differs from our studies because different time windows are used for each method, the number of samples and subjects used for each activity is different (i.e. due to utilising 63 subjects were in the supine and tripod positions and 1020 subjects were in the sitting position, the activity classification might have learned more about the sitting position, potentially biasing the results), and posture classification and biometric verification are examined separately. The number of subjects in activity classification and biometric verification differs. While having many subjects makes the study more reliable, using different people for each part reduces its consistency and reliability.

Although these CNNs are used quite frequently in the literature [19, 40], they have many differences in terms of learning methods and structures. For instance, ResNet and DenseNet CNNs possess inherent limitations. In the case of ResNet, the shortcut connection among convolutional blocks employed to stabilise training may also constrain its representational capacity. Conversely, DenseNet possesses a higher capacity due to its utilisation of multilayer feature concatenation [49]. However, this dense concatenation introduces the issue of increased memory and training time requirements [50]. In scenarios where computational resources are constrained, ResNet may be a more suitable choice than DenseNet due to its reduced memory and training time demands. DenseNet is utilised to address the vanishing gradient problem, similar to ResNet [51]. In ResNet, certain layers may provide minimal or no information, whereas, in DenseNet, information is preserved through its structure. ResNet layers have distinct weights and structures, whereas DenseNet contains cross-layer connections and a feed-forward approach [52]. This means that the results of each layer serve as inputs for subsequent layers [53]. DenseNet integrates both preserved and new information, enabling it to differentiate between the two. It boasts a higher number of feature maps compared to other architectures. Furthermore, DenseNet is effective in preventing over-fitting when working with small training sets.

CNN models are named according to the type of DL architecture and the number of blocks within the structure. For example, when we say ResNet50, we mean a ResNet CNN model with 50 residual layers. According to Lu et al. [54], the accuracy rate increases as the number of layers increases, but the training time also becomes longer because the depth of the model increases according to the number of layers. A GoogleNet (aka Inception-V1) comprises 9 inception layers and is commonly employed to preserve fine details within images. The original GoogleNet CNN has 22 deep layers. The architecture aims to achieve high accuracy while minimizing computational costs compared to previous CNN models [55]. The GoogleNet utilises filter sizes of 5 × 5, 3 × 3, and 1 × 1 to partition images of varying resolutions, thereby capturing more information from the image and addressing the issue of redundant information [55]. Therefore, GoogleNet, ResNet50 and DenseNet201 CNNs were utilised and compared in our study to see the effects on the performances.

3 Materials and methods

Our hardware experimental setup was run on an Intel(R) Core(TM) i5-6500 CPU and 16 GB RAM computer using MATLAB 2022a. GoogleNet, ResNet50 and DenseNet201 CNN models with varying time window sizes (i.e 2 s, 4 s and 10 s), different time-frequency representations, Adam (Adaptive Moment Estimation) [56], SGDM [57] optimisers and, 10 and 20 epoch sizes were used to explore DL parameters effects on activity classification. Since the amount of data we used in the training process was small, the number of epochs used was small. DL models are often trained using a combination of different datasets [40, 58] or data from various sensors [59]. This approach aims to increase the amount of training data and improve the model's performance. However, in our study, we aimed to achieve better results on both medically approved and wearable devices by using a single dataset consisting solely of ECG signals. This approach allowed us to investigate the challenge of training DL models with a limited amount of data. The general pipeline of the proposed framework is shown in Fig. 2.

Our DL parameters were adjusted based on ECG studies in the literature [19, 28, 39, 40, 60]. For instance, Thompson et al. [60] used ECG data collected from 70 subjects in the CNN model with 50 epochs, but overfitting was observed after 20 epochs. Since the number of subjects and samples in our study was less, 20 epochs and 10 epochs were used. In addition, a 60% Dropout layer was applied in the GoogleNet model and the results were observed so that the risk of overfitting was avoided. Byeon and Kwak [40] compared many DL parameters and stated that using an initial learning rate of 0.0001 and a mini-batch size of 30 achieved optimum results. In addition, they used 20 and 5 epoch to train their several DL models. Although the same configurations as Byeon and Kwak [40] study are not used, it is similar to our study because it trains many DL models with various time-frequency representations. For this reason, their model's parameters was tried in our study and setting for

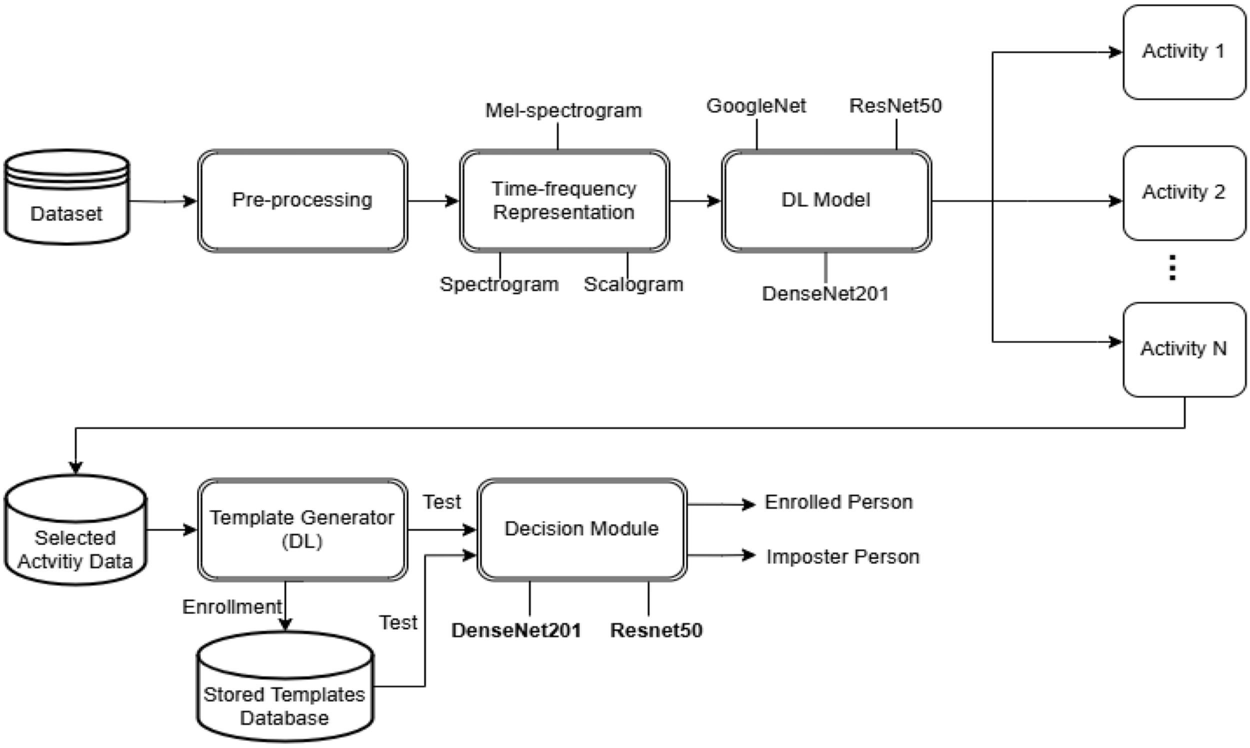

Fig. 2 The general pipeline of the proposed framework

optimum parameters for our study. According to the performances of different DL parameters in activity classification, we tuned parameters for biometric verification as 4 epochs, 0.0003 initial learning rate and 32 mini-batch size. While other parameters were kept constant for easy comparison of biometric verification and activity classification results, the number of epochs was reduced to prevent overfitting. We reduced the number of epochs to 4 in the biometric verification case because we trained our own model for biometric verification, whereas we used pre-trained structures with the transfer learning method for activity classification. Given that the number of images used in biometric verification is even smaller after activity classification, we observed that using 10 epochs led to overfitting in our trained DL models.

We explored two sets of experiments on the biometric verification framework performances prior to and after activity classification. The data utilised in this study is the Vollmer dataset [61]. This section explains the dataset, models, and experiment protocols used in the study.

3.1 Database

The Vollmer dataset [61],which is publicly available in Physionet [62], consists of data collected from 13 healthy subjects simultaneously using 5 different devices during December 2017. The individual records range from 29.18’ to 39.62’ [61].These devices include the clinically certified NeXus-10 MKII (8000 Hz) [63], eMotion Faros 360° (1000 Hz) [64], SOMNOtouch NIBP (512 Hz) [65], as well as the consumer products Hexoskin Hx1 (256 Hz) [66] and Polar RS800 Multi (1000 Hz) [8]. The Polar device is unable to measure raw ECG data. However, it can measure R-peaks, which are used as reference points. Vollmer’s study synchronised the positions of R-peaks measured by other devices with those measured by the Polar device [41]. This was done to ensure that all devices had a sampling frequency of 256 Hz and that all heartbeat locations were the same. The synchronised signals were provided in PhysioNet [62] and we used the synchronised signals.

Data was collected during four different tasks. These are resting, walking on a treadmill at a speed of 1.2 m/s, standing still, and uphill walking on the treadmill at the same speed with a 15% track inclination. Each task has at least 5 min of recordings separately.

This study compared the medically approved (CE class IIa and FDA 510 k certificates) Faros device and the wearable Hexoskin device. The sensors in the Faros [64] device consist of three attachable patches on the right chest, left chest and right abdomen. The Hexoskin [66] device is a smart shirt with textile sensors. Sensor locations are upper right and upper left abdomens and bottom left abdomen. The Faros and Hexoskin devices were selected for their comparable performance in the ML model [9], which outperformed other devices. In addition, there is no other study in which the Vollmer dataset is evaluated with DL models in terms of the proposed biometric verification framework.

3.2 Pre-processing

In the Vollmer dataset, the data collected for each subject is not equal. In our study, a total of 20 min of data, 5 min for each activity per person, was selected using the provided data labels. The sampling rate of each device was lowered from 256 Hz to 200 Hz to simplify the computation of P, Q, R, S and T points and the process of filtering and smoothing signals. Unfiltered and filtered signal sample from Subject #1 from the Faros device are shown in Fig. 3.

The Vollmer dataset contains wearable devices and attachable patches for ECG signal recordings. However, during activities, noises produced by muscles other than the heart and friction can be seen in the signal. Even though the signals provided by Vollmer et al. [41] are pre-processed using a trimmed moving average filter and Z-score normalisation, it has been observed that these filters are insufficient to remove the noise from the signal. To address this issue, a third-order Butterworth band-pass filter with 0.5 Hz low cut-off frequency and 45 Hz high cut-off frequency, along with a mean filter, were used to remove noise and provide signal smoothing. The differences between signals from the same subject and across subjects are shown in Fig. 4. The Fig. 4 illustrates two distinct samples obtained from the same subject at top, whereas the samples collected from different subjects within the same time period are shown at the bottom.

In comparison to classical ML models, DL models can produce more accurate results in a shorter time window. Shorter-time window sizes have been frequently investigated in the literature [14, 19]. Based on previous studies, the data were divided into 2-s, 4-s and 10-s time windows with a 1-s stride between each window in our study. This stride is meant to prevent overfitting and to make clear the separation between each heartbeat. In a realistic biometric verification scenario, the sample period taken from the subject and used for authentication should be short. For this reason, obtaining a low error rate even when shorter time windows are used indicates that the verification model is reliable.

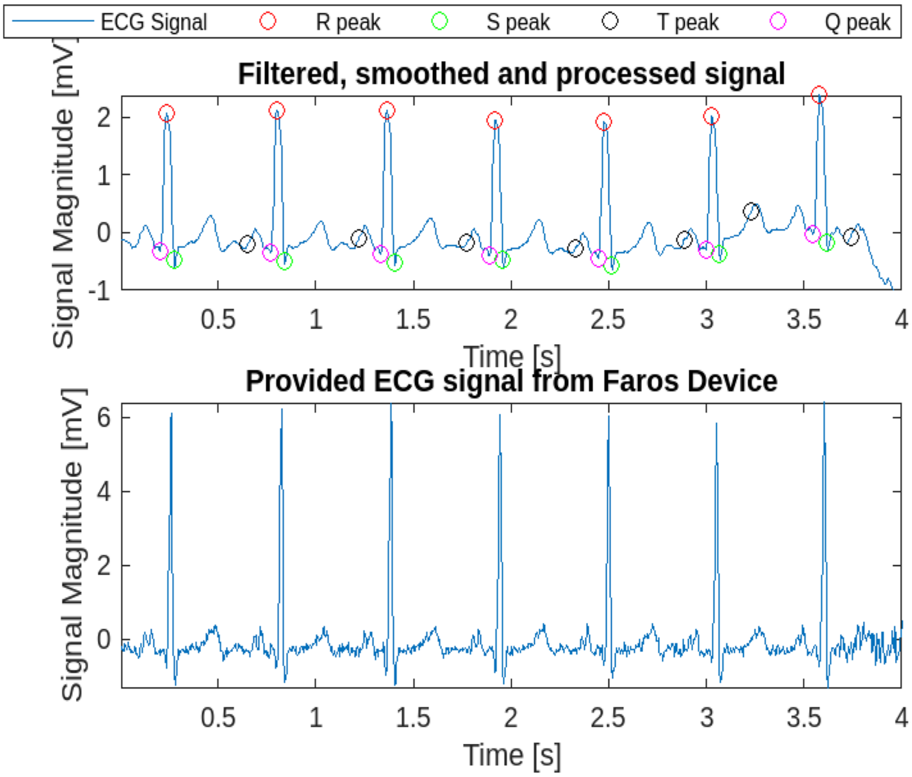

Fig. 3 A sample unfiltered and filtered signal of Subject#1 from the Faros device

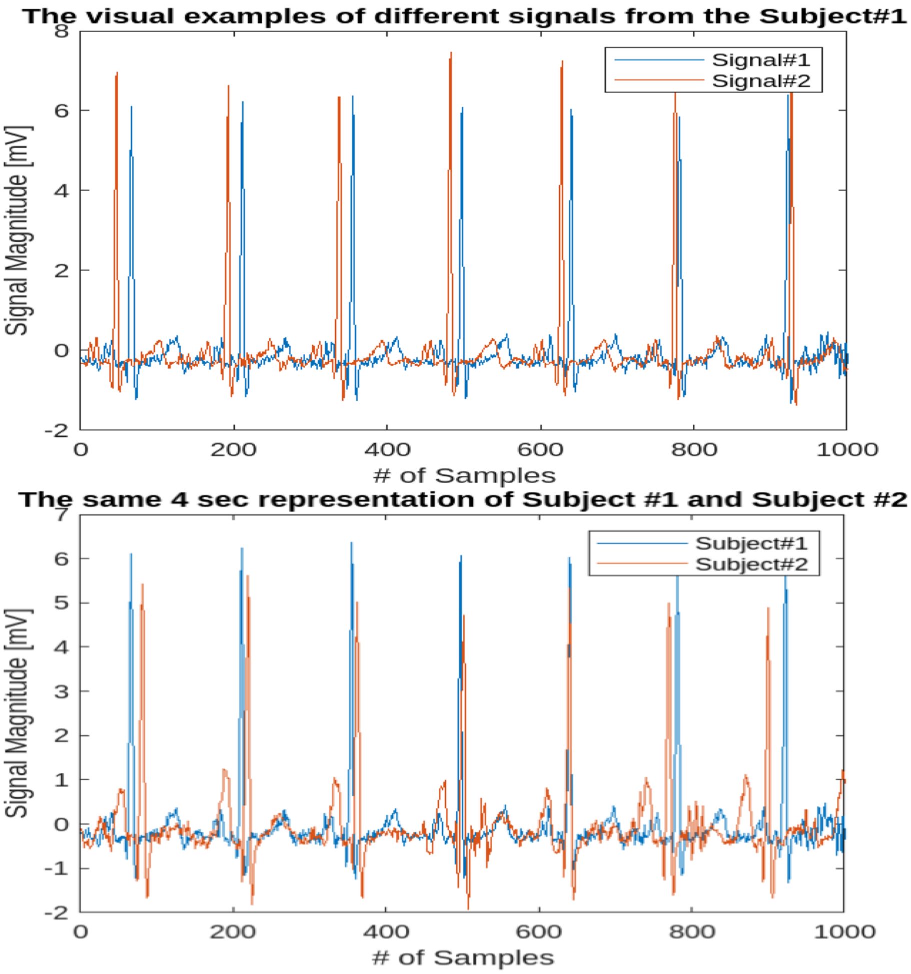

Fig. 4 The differences between signals from the same subject and across subjects

3.3 Time-frequency representations

The proposed study is the first to compare activity classification and biometric verification performances using different time windows and different time-frequency representations with different CNN structures and their parameters.

Scalogram, spectrogram and Mel-spectogram images were chosen for our study because in the literature [40], different DL models were trained with 20 epochs and 5 epochs using these images and achieved successful results. Time-frequency representations used in our study are explained in this section and they are shown in Fig. 5.

3.3.1 *Spectrogram*

A spectrogram is a visual representation of the spectrum of frequencies in a signal (e.g. sounds or ECG signals) as they vary with time [67]. In the spectrogram representation, the x-axis shows the time, the y-axis is the frequency, and the z-axis shows the energy of each frequency for a given specific time. The representation of energy is often shown with a different colour or surface in a 2D plot.

STFT of the input signals were calculated and the magnitude of the square of the STFT was used in this study to create spectrograms. The basic expression of the spectrogram is shown in Eq. (1) where S is the spectrogram, s is the signal, w is the filtering window, t represents a time axis, f represents a frequency axis and $F_s^w(t,f)$ is the STFT [68].

$$S_s^w(t,f) = \left|F_s^w(t,f)\right|^2 \tag{1}$$

Each window (2 s, 4 s or 10 s) from the pre-processed matrix is used to create a spectrogram image using a *spectrogram* toolbox in MATLAB. In the spectrograms, the interval where the signal changes more is shown in brighter colours (yellow), while the interval where the signal changes less is shown in darker colours (blue/

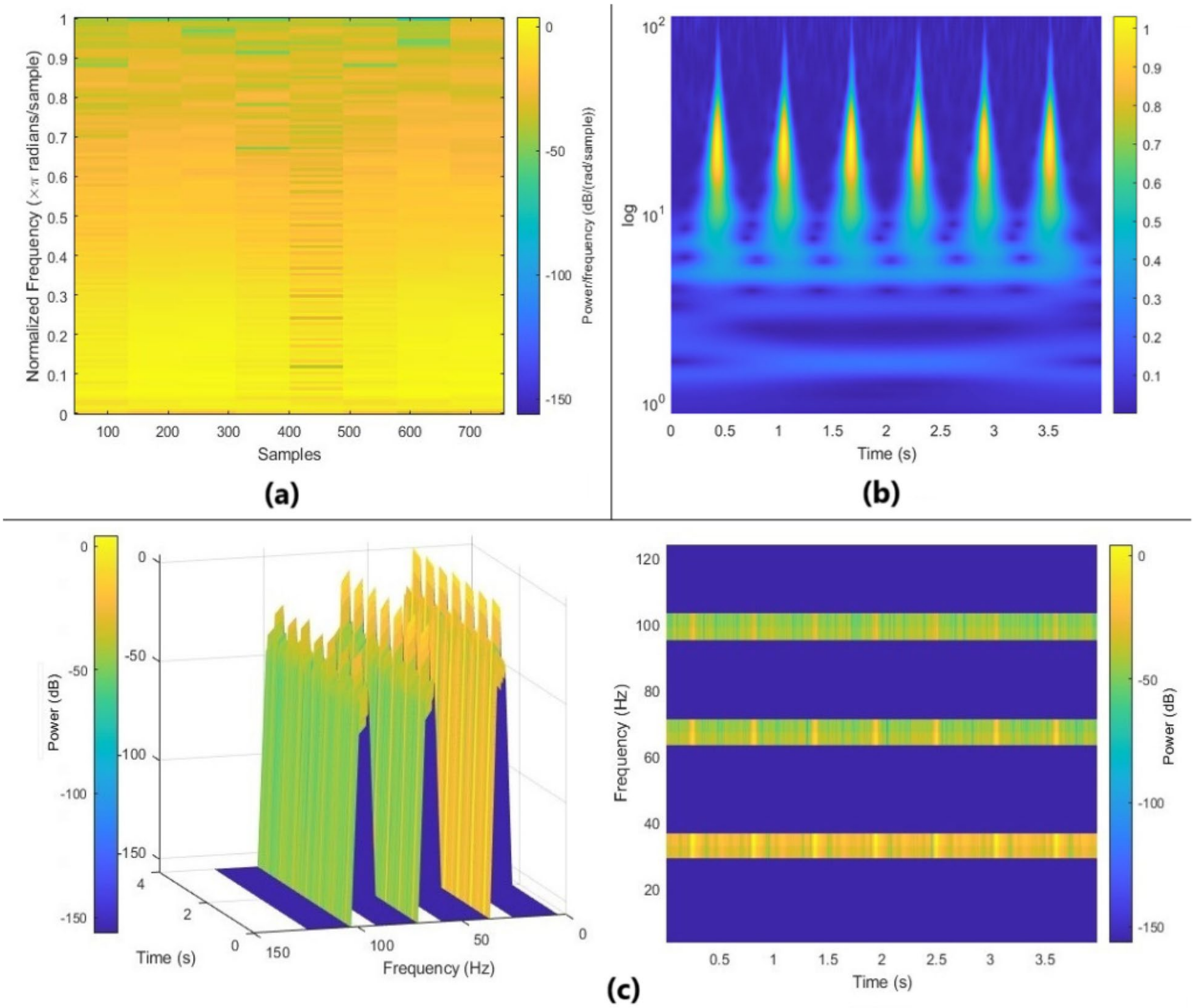

Fig. 5 A sample 4-s time windowed image of Subject #1 from the Hexoskin device. **a** spectrogram, (**b**) scalogram, (**c**) Mel-spectrogram

navy blue) [69]. Figure 5a is the sample representation of a spectrogram image.

3.3.2 Scalogram

A scalogram is a visual representation of a wavelet transform with time, scale, and coefficient axes, unlike the spectrogram, which is a visual representation of the spectrum of a time-varying signal. A scalogram is computed by obtaining the absolute value of the CWT of the signals [70]. A scalogram can be expressed as time and frequency functions. It is better suited than the spectrogram for signals that have multiple scales of features. In other words, these signals have slowly varying events that are interrupted by sudden changes such as ECG, earthquakes and audio signals. The mathematical expression of the CWT is shown in Eq. (2) [71].

$$CWT(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} f(t) * \psi\left(\frac{t-b}{a}\right) dt \qquad (2)$$

$$a \in R^{+} - \{0\}, \quad b \in R$$

In Eq. (2), $\psi(t)$ is known as the mother wavelet, while the parameters (i.e. shifting and scaling) derived from it are known as the daughter wavelet. *f(t)* represents a function, a represents the scaling factor, b represents the shifting factor and R represents *Real Numbers* [71]. New parameters were adjusted as the sampling frequency was 256 Hz and the voices per octave were 12 to obtain more precise scalogram images in a CWT filter bank [72]. Figure 5b is the sample representation of a scalogram image.

3.3.3 Mel-spectrogram

A Mel-spectrogram is a spectrogram transferred to the Mel-scale. A Mel-scale which is widely used in voice analysis accentuates the low-band frequency in voice and eliminates the high-band frequency noise [40]. The mathematical formula of the Mel-scale is shown in Eq. (3) [40, 73]. In this equation, f represents the frequency (Hz) and m represents the Mel-scale.

$$m = 2595\,\mathbf{log}_{10}\left(1 + \frac{f}{700}\right) \tag{3}$$

ECG involves substantial information which is commonly used in ECG applications at the low-band frequency [40, 74]. For this reason, the Mel-spectrogram is investigated with other time-frequency representations in our experiments.

The *melSpectrogram* toolbox in MATLAB was used to obtain the Mel-spectrogram [75]. A Mel-spectrogram was generated by applying a frequency domain filter bank to ECG signals windowed over time. This filter bank contains many band-pass filters. The centre frequency of the filter is in Hz and the time instants for each window are in seconds. The colour intensity represents the amplitude of a frequency at a certain point in time in terms of dB.

Figure 5c shows a Mel-spectrogram sample image with a 3D representation and its 90° view from above at the 256 Hz sampling frequency. Each window from the pre-processed matrix is used to create Mel-spectrogram images with 256 Hz sampling frequency. Navy blue bands (non-informative parts) appear in Mel-spectrogram images created from windows of a pre-processed matrix, regardless of the sampling frequency. The main reason for choosing a 2D image instead of a 3D image, which contains more information, is to be able to examine the image in time and frequency axes, as in scalogram and spectrogram images. It also reduces the time spent on CNN model training. When applying the Fourier transform (FT) in the transition from the time domain to the frequency domain, half the maximum frequency (f = *Fs/2*) should be applied [76]. Still, when this is applied, only a single stripe appears (at approximately 30 Hz). As '*f*' increases, the size and weight of the filters in the Mel-filter bank will change and more stripes will be visible as the size of the Mel-scale will increase, but the resolution of the image will decrease. Since the single stripe image contains insufficient information for the DL model, the sampling frequency was increased.

3.4 Classification models

This study explored the performance of different CNN models, any DL hyperparameters such as epoch sizes, time window sizes, optimisers, several enrollment samples and time-frequency representations for two different ECG recorders. This study examines two distinct verification processes: one that incorporates activity classification in biometric verification, and another that conducts biometric verification directly without categorising activities beforehand. GoogleNet, ResNet50 and DenseNet201 CNN models,which are pre-trained with the ImageNet dataset [77], were used for activity classification with transfer learning method. After examining the performance of different parameters, the ResNet50 and DenseNet201 models, which achieved successful results in activity classification, were selected and their parameters were tuned for biometric verification.

To ensure our model's independence from specific datasets and its applicability to other datasets, we utilised pre-trained CNN models for activity classification. For the biometric verification model, we trained it from scratch using data from 11 subjects collected with the Faros device. Using biometric identification data to train a DL model and use the model in biometric verification, which is a commonly used technique in the literature [78, 79], may not be sufficient alone for high confidence and reliability in the case of a small dataset. A larger and more diverse data set may improve generalisation and robustness. However, evaluating the model with metrics like FAR, FRR, EER, and accuracy rates provides insights tested on diverse data and periodically re-evaluated is crucial for maintaining its reliability and trustworthiness. For th into performance. Low FAR, FRR, and EER, along with high accuracy, indicate a reliable model. In addition, ensuring the model isis reason, this model was then validated and tested with data from the Hexoskin device, achieving high success rates across different devices. The used CNN models and their parameters will be considered in detail in the following section.

4 Experiments and results

In biometric verification tasks, time windows of 2 s, 4 s and 10 s were examined for scalogram, spectrogram and Mel-spectrogram images. The number of genuine samples used in our study is 1, 3 and 5 in the enrollment stage. If we express these samples as enrollment time, the minimum is 2 s (1 sample × 2-s time window) and the maximum is 50 s.

In the activity classification task, 72% of the images obtained were used in training, 8% in validation and 20% in testing. Results are divided into three subsections: direct biometric verification across activities, the overall accuracy of activity classification and biometric verification following activity classification.

4.1 ECG biometric verification across all activities

The data from each device has been individually examined. In each device, data were split into three parts:

training, validation and testing. For training and validation data, the data of 11 subjects were selected. Selected subjects' data were split into 80% for training and 20% for validation randomly. The remaining unseen 2 subjects were used for testing. Table 3 describes the number of images in 2, 4 and 10-s time windows utilised for training, validation and testing for both CNN structures.

In the process of training a CNN, two critical hyperparameters that can influence both the training process and the ultimate performance of the model are the mini-batch size and the initial learning rate [80]. To find optimal parameters some trials were made using similar studies [40, 80] in our ResNet50 model. After these trials, the most suitable parameters were determined with the mini-batch size set to 32, the maximum epochs of 4, the initial learning rate of 0.0003, and the execution environment set to CPU. GPUs are often preferred in large-scale DL models because of their faster processing capacity [53], but CPUs are less costly and accessible in small-scale projects. In addition, CPU use is more common in mobile devices and gives better results [81]. For these reasons, CPU was preferred in our study.

Table 3 The number of images in training, validation and testing sets for direct biometric verification

	2 sec	4 sec	10 sec
# of images in training	4994	2992	1342
# of images in validation	1254	748	330
# of images in testing	682	408	182

The ResNet50 CNN model created and used for training and biometric verification in this study is shown in Fig. 6 [82, 83]. The output size at each stage is indicated as "$n \times n$" after each block. The output of each stage is used as the input of the next stage. The model was trained to perform biometric identification, with Output1 representing the biometric identification accuracy rate. In all cases, training and validation accuracy rates were higher than 90% (i.e. minimum 90%). For this reason, the number of epochs has not been increased to avoid over-fitting.

Figure 7 illustrates the details of the DenseNet201 CNN model used for training and biometric verification. Certain structures have been abbreviated and represented with coloured blocks. An "Output 1" shows the accuracy of the training, while "Output 2" represents the result of the biometric verification in terms of EERs.

4.1.1 General procedures of training and validation

The 11 subjects allocated for training and validation trained the CNN model as if it were the identification task, while the 2 unseen subjects were used for biometric verification. To increase the reliability of the model, we used different subject pairs as unseen subjects in each testing stage. Unseen subject pairs were selected as *P1-P2*, *P6-P7*, *P8-P9* and *P11-P12*. The system is retrained

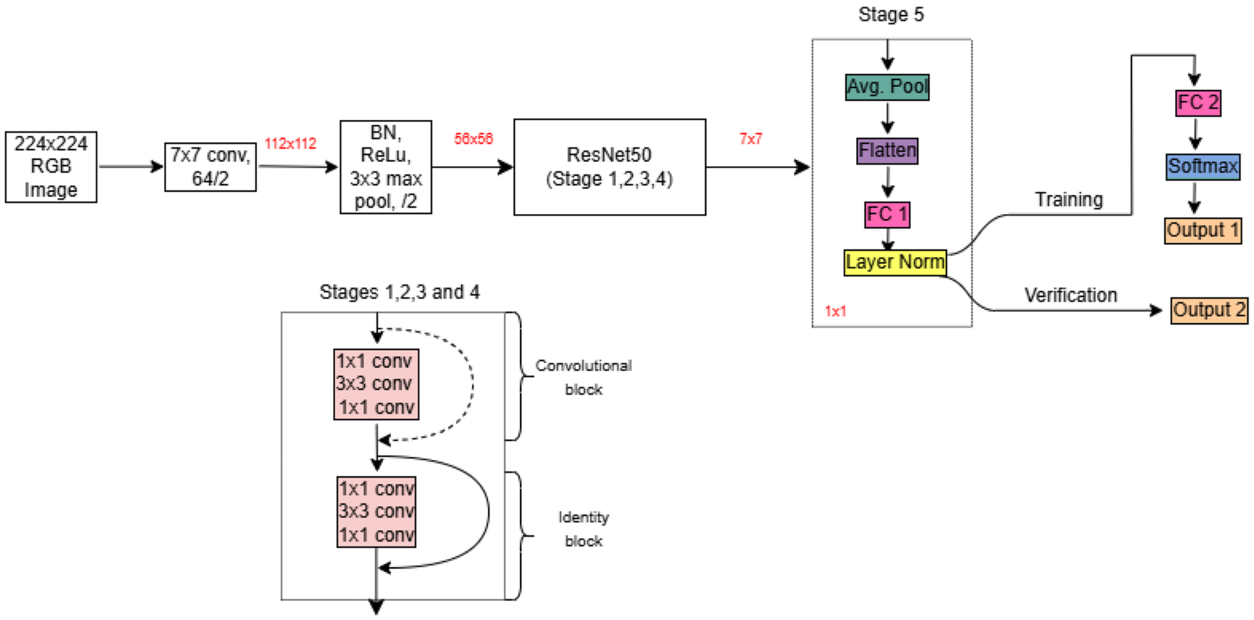

Fig. 6 Used ResNet50 model in biometric verification

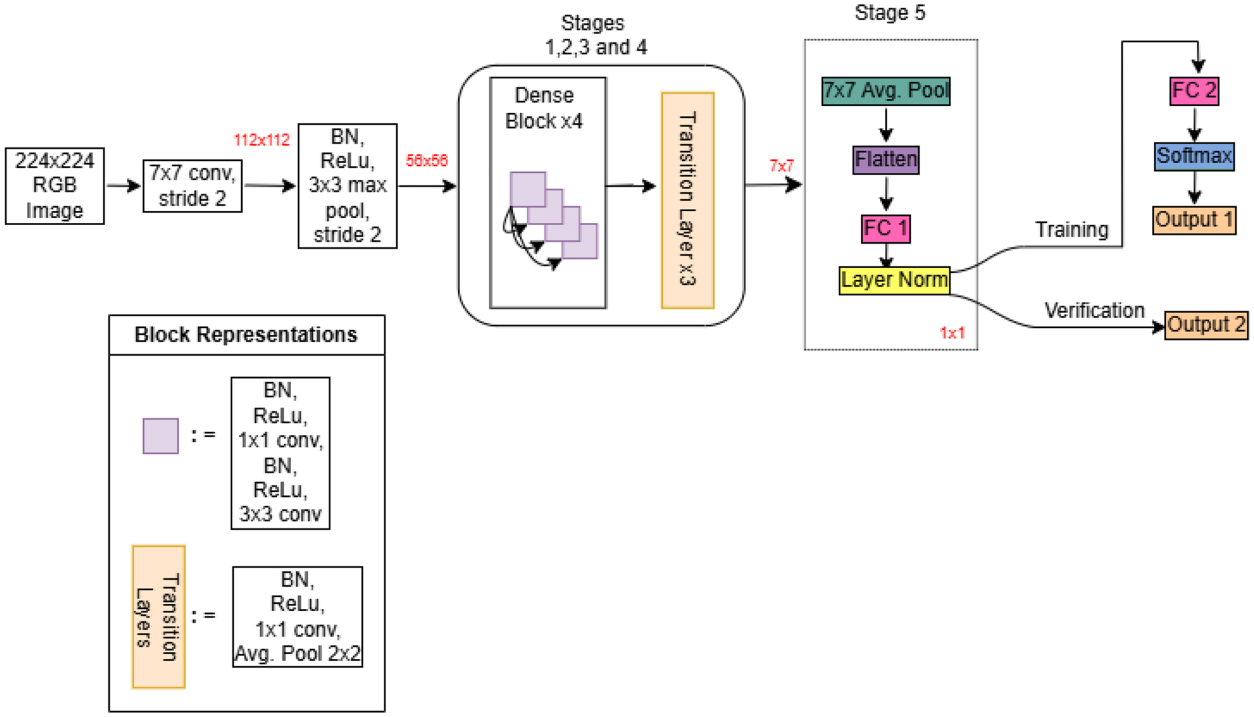

Fig. 7 Used DenseNet201 model in biometric verification

for each pair of subjects, as the samples from the remaining 11 subjects will vary for each pair of unseen subjects.

To increase reliability, cross-validation is another technique in DL for evaluating model generalisation, tuning hyperparameters, understanding the bias-variance tradeoff, ensuring robustness, and maximising data utilisation [84]. It involves splitting the dataset into multiple folds to provide a reliable performance estimate, select optimal hyperparameters, and ensure the model's performance is not dependent on a specific train-test split. This technique helps to make the most of the available data and improve the overall effectiveness of the model. In this study, k-fold cross-validation ($k = 5$) was used in the Sc4 biometric verification scenario. When the P6-P7 pair was utilised as unseen subjects, the standard deviation of the validation accuracy rates was 0.0212, compared to 0.0216 for other pairs. To reduce computational-costs, subsequent experiments were conducted using randomly separated training, test, and validation sets instead of cross-validation. The *Flatten, FC1* (fully connected (FC)) and *layer normalisation* layers are used to produce templates for biometric verification. To enable the system to function as a classification problem during the training phase, the *FC2, Softmax* and *Output1* (classification output) layers were added. In testing, genuine images from one unseen subject, alongside another imposter subject were used. Each selected image was used to create the verification embedding. The genuine and imposter samples were balanced and randomly selected. Due to the challenges associated with determining an appropriate training and testing ratio, we have elected to utilise 1, 3 and 5 genuine samples for each time window condition.

After the validation was completed, we saved this trained model, which is called directed acyclic graph network (DAGNetwork) [85, 86], as an embedded model and tested this model with the data of unseen subjects. This *DAGNetwork* is subsequently saved and the final three layers are removed. As depicted in Figs. 6 and 7, the design enables output from the "Layer norm" layer. This design is saved as a second *DAGNetwork* for testing. It is used for making predictions on testing data via the "predict" toolbox in MATLAB R2022a. In this manner, our embedded model is successfully constructed. The embedded model is subjected to a verification task wherein its performance is evaluated on previously unseen subjects. For each subject, a predetermined number (N= 1, 3 or 5) of images are utilised for enrollment and subsequently input into the trained model to generate a user template. Templates for each subject are calculated by the average of the embeddings which were obtained from N genuine samples from the same subjects. This relationship can be mathematically represented as depicted in Eq. (4). In this formula, $(X_i)_s$ represents i_{th} enrollment image of the selected subject s. M symbolises the created embedding model while T_s is the template for the selected subject [87].

$$T_s = \frac{1}{N}\sum_{i=1}^{N} M((X_i)_s) \tag{4}$$

The number of N samples for templates is randomly selected from each person's embeddings. Subsequent to this procedure, imposter samples and genuine samples from both the same subject class and different classes within the testing set are introduced to the model. The Euclidean distances between the verification embeddings and the user templates are then computed. If the distance is below a predetermined threshold, the verification response is deemed positive; otherwise, it is negative. The threshold is determined by evaluating the FAR and FRR and selecting the threshold that minimises the combined error. Alternatively, it maximises the AUC in the corresponding ROC.

Mean EERs are shown for different numbers of genuine samples and three image representations in Table 4. *Sc2, Sc4* and *Sc10* are scalogram images that were created from 2-s, 4-s and 10-s time windows. *Sp2, Sp4* and *Sp10* are spectrogram images and *Mel2*, *Mel4* and *Mel10* represent Mel-spectrogram images with the same time windows. *1*, *3* and *5* are genuine samples used in enrollment.

A general trend was observed in which the EER decreased with an increase in the number of genuine samples. Furthermore, in ResNet50, the biometric verification performance of the Faros medical device outperforms that of the consumer-based Hexoskin device. However, in DenseNet201, the opposite is observed. The Hexoskin device, being wearable, may be more sensitive to body movements and natural signal variations. This sensitivity could result in ECG signals that are better suited to the DenseNet201 model. The close fit and ability to track the body might help Hexoskin capture specific signal features that DenseNet201 can effectively leverage. DenseNet201's dense connectivity between layers may aid in learning complex, variable patterns in the data.

Scalogram images generally yielded the lowest EERs, while Mel-spectrograms obtained the highest EERs. In the DenseNet201 case, although Mel-spectrograms generally exhibited the highest EERs, their performance was comparable to other representations within a 2-s time window for both devices. An increase in the time window length (from 2 to 10 s) for spectrogram and Mel-spectrogram images resulted in lower EERs, whereas the minimum EERs for scalogram images were obtained with a 4-s time window. It is expected that a longer enrollment time will give a lower error rate, but the best results in the *Sc4* case indicate that the information from the *Sc10* case causes errors in the DL model or that the *Sc4* pictures present more distinctive features.

4.2 Activity classification

After the images were separated for training, verification and testing with the rates specified in Sect. 4, DL parameters were adjusted in accordance with the biometric verification model. However, since it will be classified in a 4-class activity classification, the number of epochs was increased and the effects of optimisers on the classification were examined. Table 5 describes the number of images in 2, 4 and 10-s time windows utilised for training, validation and testing for all CNN structures in activity classification.

The training process was optimised by adjusting various parameters. The initial learning rate, set at 0.0003, determined the initial step size towards the negative gradient of the loss function. The mini-batch size, set at

Table 4 Mean EERs (%) are shown for different numbers of genuine samples and three image representations

	ResNet50						**DenseNet201**					
Devices	**Faros**			**Hexoskin**			**Faros**			**Hexoskin**		
# of genuine samples	**1**	**3**	**5**	**1**	**3**	**5**	**1**	**3**	**5**	**1**	**3**	**5**
Image types												
Sc2	37.57%	33.28%	32.95%	41.48%	36.45%	34.46%	36.55%	33.58%	34.20%	35.78%	31.12%	30.79%
Sc4	**15.93%**	**13.42%**	**11.15%**	34.01%	**12.19%**	**11.21%**	22.98%	**16.85%**	**12.81%**	22.49%	**10.60%**	**8.88%**
Sc10	22.39%	19.78%	18.96%	**22.53%**	20.47%	19.09%	**22.25%**	20.74%	20.88%	20.47%	19.92%	18.13%
Sp2	38.09%	38.53%	37.35%	39.11%	33.28%	34.64%	39.48%	38.56%	37.35%	38.78%	36.58%	35.30%
Sp4	34.97%	30.64%	28.62%	35.97%	30.64%	28.62%	40.81%	36.52%	36.21%	33.95%	29.29%	27.76%
Sp10	27.06%	26.83%	26.96%	24.39%	20.95%	21.23%	26.65%	26.79%	26.24%	**19.58%**	19.03%	18.20%
Mel2	38.71%	37.68%	38.67%	38.34%	39.22%	39.48%	36.44%	36.58%	38.16%	39.26%	38.93%	41.06%
Mel4	39.64%	37.99%	37.75%	35.50%	34.27%	35.68%	38.97%	35.60%	35.78%	38.56%	37.21%	35.62%
Mel10	32.83%	32.69%	30.08%	28.10%	28.79%	27.55%	33.10%	35.16%	34.34%	29.17%	28.90%	26.85%

Table 5 The number of images in training, validation and testing sets for activity classification

	2 sec	4 sec	10 sec
# of images in training	5316	3184	1426
# of images in validation	592	352	157
# of images in testing	1476	884	396

32, specified the subset of the training set used in each iteration. The maximum number of epochs, assigned at either 10 or 20, determined the maximum number of full passes of the training algorithm over the entire training set. Selecting an appropriate number of epochs is crucial; too few can result in under-fitting while too many can lead to over-fitting. Additionally, two optimisers, Adam and SGDM, were compared. These optimisers serve to update the network's weights during training to minimize the loss function, with the selection of optimizer having a significant impact on model performance [88]. Early-stopping techniques can be used to automatically determine the number of epochs and prevent overfitting [89]. However, they may not be useful due to the risk of stopping too early and being sensitive to the validation set [89]. In our study, the manual method was used because the validation set was small and the hyperparameters had to be similar to those of other cases.

In MATLAB, the GoogleNet, ResNet50 and DenseNet201 CNNs, which have been pre-trained on the ImageNet dataset to classify 1000 classes, are available to users. In this experiment, rather than training the models from scratch, pre-existing models, pre-trained on the extensive ImageNet dataset, were utilised. In Fig. 8, the layers indicated in colour have been incorporated for the purpose of 4-class classifications in all CNN models. The remaining layers are in the original GoogleNet, ResNet50 and DenseNet201 architectures. Figure 8 illustrates the fundamental structure of GoogleNet (a), ResNet50 (b) and DenseNet201 (c) CNN models in activity classification.

In GoogleNet, to prevent overfitting, it was necessary to include a dropout layer. An *FC* layer was adjusted to classify 4 activity classes. The *Output* layer indicates the activity classification of the evaluated time-frequency representation images. A mean accuracy rate obtained from the GoogleNet, ResNet50 and DenseNet201 CNN models for activity classification is shown in Table 6. The highest accuracy rates for each device are denoted in bold.

According to GoogleNet results, the model which was used with Hexoskin data had higher accuracy rates for spectrogram images while the model with the Faros data had better performance for scalogram and Mel-spectrogram images. Limiting training epochs to 20 shows minimal improvement in accuracy beyond this point. Both optimisation algorithms perform similarly when applied to the GoogleNet CNN. Faros device performs better with SGDM while Hexoskin performs better with Adam optimiser.

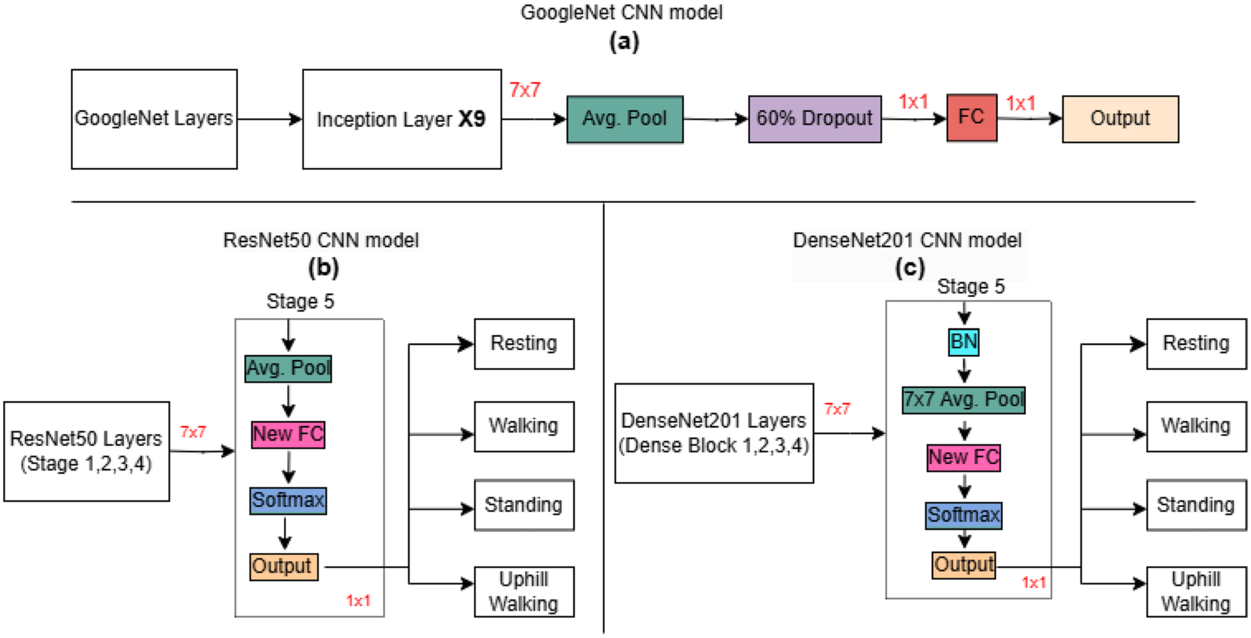

Fig. 8 CNN models in activity classification

Table 6 Accuracy rates (%) in activity classification across four individual activities

		FAROS				HEXOSKIN			
		10 ADAM	**20 ADAM**	**10 SGDM**	**20 SGDM**	**10 ADAM**	**20 ADAM**	**10 SGDM**	**20 SGDM**
GoogleNet	Sc2	61.80%	68.50%	63.08%	65.04%	59.01%	**63.55%**	58.74%	62.53%
	Sc4	56.79%	69.57%	60.52%	**70.81%**	60.29%	61.88%	59.73%	56.79%
	Sc10	41.41%	53.28%	52.27%	54.04%	49.50%	50.76%	46.21%	56.31%
	Sp2	50.34%	**53.93%**	48.58%	51.90%	50.41%	50.75%	49.66%	48.85%
	Sp4	50.34%	49.55%	50.57%	53.62%	53.17%	54.19%	52.49%	**55.20%**
	Sp10	41.16%	49.24%	46.72%	42.68%	50.51%	53.54%	45.20%	40.91%
	Mel2	25.00%	**52.58%**	46.95%	50.81%	25.00%	25.00%	37.20%	**49.73%**
	Mel4	41.52%	50.23%	47.62%	44.57%	25.00%	25.00%	41.29%	46.15%
	Mel10	33.59%	25.00%	44.70%	51.01%	25.00%	25.00%	33.33%	41.67%
ResNet50	Sc2	65.58%	67.28%	66.13%	65.31%	61.25%	64.91%	**65.18%**	61.31%
	Sc4	**68.55%**	67.20%	61.20%	64.48%	62.56%	62.33%	64.03%	64.14%
	Sc10	55.05%	44.70%	53.03%	51.26%	58.08%	56.31%	57.32%	60.35%
	Sp2	55.01%	55.96%	52.29%	52.78%	51.42%	**54.74%**	50.07%	49.53%
	Sp4	58.60%	**61.43%**	56.90%	57.47%	54.64%	51.58%	51.70%	52.83%
	Sp10	48.49%	53.28%	49.75%	53.28%	46.97%	47.48%	49.24%	54.29%
	Mel2	59.35%	57.59%	57.32%	58.33%	**59.76%**	58.13%	53.39%	53.05%
	Mel4	56.90%	56.34%	60.52%	**61.09%**	57.35%	58.26%	50.68%	59.62%
	Mel10	56.06%	45.96%	55.81%	59.60%	49.50%	47.73%	51.26%	50.00%
DenseNet201	Sc2	68.29%	**69.85%**	68.16%	65.31%	64.43%	64.23%	**65.79%**	65.11%
	Sc4	65.16%	69.34%	65.84%	63.69%	65.05%	62.44%	65.16%	63.91%
	Sc10	55.30%	50.25%	49.75%	54.29%	53.79%	59.34%	57.07%	58.84%
	Sp2	56.57%	**60.43%**	52.98%	55.83%	54.07%	**55.69%**	51.36%	54.00%
	Sp4	53.62%	57.47%	54.30%	54.64%	55.66%	54.64%	52.49%	53.51%
	Sp10	47.73%	58.08%	51.52%	54.55%	47.73%	53.54%	49.24%	53.03%
	Mel2	60.84%	59.15%	60.37%	62.13%	56.98%	56.44%	55.42%	56.64%
	Mel4	63.01%	61.09%	58.60%	**65.50%**	54.07%	53.28%	57.69%	**58.24%**
	Mel10	47.98%	50.76%	57.58%	58.33%	46.21%	43.94%	51.01%	55.30%

ResNet50 results show the Faros device had better classification outcomes than the Hexoskin device. Models with 20 epochs and SGDM optimisers in 10-s time windows were more successful than those with 10 epochs and Adam optimisers. It is unclear if one optimiser is better than the other. Using scalogram images with 10 Adam produced better accuracy rates, especially with the Faros device. The maximum accuracy rate was achieved with the 10 Adam scenario using scalogram images with a 4-s time window.

DenseNet201 accuracy rates show that the Faros device had higher accuracy rates than the Hexoskin device. The Faros device performed best with 20 Adam for scalogram and spectrogram images, and 20 SGDM for Mel-spectrogram images. The Hexoskin device did not show a clear difference between 10 and 20 epochs or between the optimizers.

4.3 Activity-aware biometric verification

As a general framework, we aimed to classify activities and create biometric verification models for each activity in DL models. We classified activities using different CNN models, but could not achieve high accuracy rates in activity classification through DL models. There could be several factors contributing to this, such as the inadequate training of the system and the images selected not being distinct enough for the activities.

In DL models, utilising biometric verification with the newly assigned class label, regardless of whether the samples are classified incorrectly or correctly, leads to more errors and an unrealistic scenario. This is because of the insufficient accuracy rates obtained from DL models in activity classification and the negative impact of excessive errors in activity classes on our ability to analyse the influence of each activity on biometric verification.

Table 7 The number of images in training, validation and testing sets for each activity [61]

	2 s	**4 s**	**10 s**
# of images in training	1254	748	330
# of images in validation	308	187	88
# of images in testing	170	102	46

Therefore, we manually divided time-frequency representations into real activity classes and used ResNet50 and DenseNet201 CNN models for biometric verification. Since this study is a follow-up phase of Sect. 4.1, all hyperparameters and procedures (see Figs. 6 and 7) were identical in CNN models. Table 7 includes information on the number of images used for training, validation and testing per activity in 2, 4 and 10 s time windows for both CNN structures.

This study is important to see if using activity classification can improve the direct biometric verification model from Sect. 4.1. Moreover, the study aims to investigate the impact of different activities on biometric verification. Table 7 shows a decrease in the number of training images as the time window size increases. However, the number of heartbeats in the images increases correspondingly. As the enrollment time increases, the number of images in the training set of the DL model decreases, which affects the EERs. Whether these are sufficient for biometric verification can be evaluated by examining the EERs. Since testing the model for only one user would be insufficient to evaluate the reliability of the model, we tested the model for the same pairs of subjects as described in Sect. 4.1.

Table 8 shows the EERs of the two devices tested using the ResNet50 CNN model while Table 9 shows the EERs of the DenseNet201 model for each activity.

For both CNN models, when analysing the effectiveness of time-frequency representations, it was found that the scalogram had the lowest EERs, followed by the spectrogram in the second place, and the Mel-spectrogram in the last place. The optimal results for scalogram images are achieved with 4-s time windows, followed by 10-s time windows, and then 2-s time windows. Furthermore, cases containing *3* and *5* genuine samples generally outperformed *1* sample cases while containing similar EER results. For both devices, activities that involved less movement (i.e. resting and standing) generally achieved more successful results than those that involved more movement (i.e. walking and uphill walking). The 10-s time windows for Mel-spectrogram and spectrogram images yield the lowest EERs, followed by the 4-s time windows and then the 2-s time windows cases. EER results alone are not sufficient to evaluate the reliability of the biometric model. It is also necessary to mention the rates of samples that the model incorrectly accepts (FAR) or incorrectly rejects (FRR). In a reliable model,

Table 8 Biometric verification performances of ResNet50 CNN model in terms of EER for each activity

	Activities	**Resting**			**Walking**			**Standing**			**Uphill walking**		
Devices	**# of genuine samples**	**1**	**3**	**5**	**1**	**3**	**5**	**1**	**3**	**5**	**1**	**3**	**5**
	Image types												
Faros	Sc2	19.12%	17.21%	20.00%	28.97%	28.38%	28.38%	28.38%	21.76%	22.50%	23.38%	30.59%	30.59%
	Sc4	6.99%	7.23%	**6.25%**	5.88%	**4.17%**	**4.17%**	7.60%	**5.88%**	6.13%	11.52%	11.03%	**9.80%**
	Sc10	12.62%	10.45%	10.99%	16.97%	13.71%	13.71%	13.17%	12.62%	10.99%	20.77%	19.27%	18.60%
	Sp2	24.56%	23.53%	25.15%	31.91%	31.62%	29.12%	30.15%	26.47%	26.03%	34.12%	32.94%	33.53%
	Sp4	28.43%	28.92%	26.72%	26.72%	23.53%	24.75%	30.15%	27.94%	28.43%	37.44%	34.07%	33.64%
	Sp10	17.51%	16.97%	**16.43%**	19.14%	12.62%	**10.45%**	15.34%	**14.80%**	**14.80%**	25.67%	**22.60%**	**22.60%**
	Mel2	28.53%	30.74%	29.26%	36.03%	37.21%	37.06%	32.35%	33.38%	34.41%	35.15%	35.88%	36.91%
	Mel4	27.70%	26.72%	28.43%	29.17%	29.17%	30.15%	35.29%	33.82%	35.05%	33.82%	34.31%	36.27%
	Mel10	19.57%	**18.48%**	**18.48%**	21.27%	16.85%	**15.84%**	23.91%	**23.37%**	23.91%	32.07%	**29.35%**	**29.35%**
Hexoskin	Sc2	25.44%	21.56%	21.98%	26.08%	26.55%	24.20%	26.55%	24.79%	22.86%	25.64%	24.35%	22.22%
	Sc4	0.98%	0.98%	**0.74%**	4.29%	4.04%	**3.80%**	0.61%	0.61%	**0.49%**	8.24%	**6.72%**	7.43%
	Sc10	15.83%	13.12%	12.03%	16.38%	9.86%	15.97%	14.13%	12.50%	9.24%	22.28%	22.28%	18.34%
	Sp2	30.82%	26.90%	26.47%	33.51%	31.35%	30.96%	33.09%	31.32%	31.03%	38.25%	36.40%	35.76%
	Sp4	25.74%	23.53%	21.57%	17.40%	16.42%	16.24%	22.79%	21.65%	19.54%	29.06%	26.65%	25.61%
	Sp10	12.62%	10.99%	**8.82%**	16.43%	12.62%	**10.99%**	11.75%	9.29%	**8.27%**	23.71%	21.54%	**19.99%**
	Mel2	38.28%	36.47%	35.44%	37.65%	36.62%	36.47%	37.35%	36.32%	35.44%	39.44%	37.65%	37.80%
	Mel4	31.35%	31.63%	31.37%	30.92%	31.67%	32.51%	34.88%	33.57%	33.29%	33.08%	34.18%	34.81%
	Mel10	25.54%	23.91%	**23.12%**	23.71%	**21.47%**	21.74%	23.37%	21.20%	**20.11%**	28.00%	**25.66%**	25.72%

Table 9 Biometric verification performances of DenseNet201 CNN model in terms of EER for each activity

	Activities	Resting			Walking			Standing			Uphill walking		
Devices	**# of genuine samples**	**1**	**3**	**5**	**1**	**3**	**5**	**1**	**3**	**5**	**1**	**3**	**5**
	Image types												
Faros	Sc2	20.44%	19.26%	19.56%	27.65%	28.38%	27.21%	21.18%	23.53%	23.53%	29.56%	31.91%	30.44%
	Sc4	8.21%	7.72%	**6.74%**	5.39%	5.15%	**4.29%**	**4.41%**	5.15%	4.66%	11.03%	12.50%	**10.78%**
	Sc10	11.96%	10.33%	9.24%	17.39%	14.67%	13.59%	15.76%	13.17%	12.08%	18.06%	20.23%	19.14%
	Sp2	25.15%	26.32%	24.41%	33.97%	35.15%	32.50%	31.88%	28.38%	27.65%	36.03%	35.74%	35.44%
	Sp4	23.28%	23.04%	25.49%	23.28%	26.23%	25.74%	25.98%	29.17%	26.72%	35.42%	29.72%	28.92%
	Sp10	18.60%	18.06%	**16.43%**	19.69%	12.08%	**10.45%**	17.51%	18.60%	**16.43%**	25.12%	**24.58%**	25.67%
	Mel2	29.41%	30.29%	30.00%	36.18%	36.40%	37.35%	33.53%	33.97%	32.65%	34.56%	35.44%	35.44%
	Mel4	31.62%	31.62%	32.84%	30.64%	31.86%	29.90%	35.05%	35.54%	35.76%	34.56%	35.05%	35.54%
	Mel10	20.11%	**18.48%**	20.11%	**17.39%**	18.48%	21.74%	24.46%	22.28%	**21.74%**	**27.72%**	30.98%	29.35%
Hexoskin	Sc2	19.12%	15.59%	15.59%	27.65%	25.59%	22.50%	20.00%	17.79%	17.21%	24.56%	24.41%	22.79%
	Sc4	**1.96%**	2.94%	2.94%	3.19%	**2.21%**	2.45%	**0.49%**	**0.49%**	**0.49%**	5.64%	6.62%	6.86%
	Sc10	13.05%	11.41%	10.87%	10.06%	9.51%	8.42%	9.51%	8.42%	6.79%	19.57%	17.93%	17.96%
	Sp2	29.71%	26.18%	25.74%	28.09%	28.09%	25.59%	27.94%	28.68%	28.98%	31.91%	30.59%	31.18%
	Sp4	22.79%	22.79%	20.10%	16.91%	14.46%	14.71%	20.34%	20.10%	18.63%	27.66%	28.74%	27.65%
	Sp10	8.82%	8.27%	**7.73%**	8.82%	6.64%	**6.10%**	7.73%	7.19%	**5.56%**	21.86%	**20.23%**	20.77%
	Mel2	35.59%	35.59%	35.91%	35.88%	35.74%	36.03%	38.82%	35.59%	36.62%	39.01%	38.24%	38.97%
	Mel4	32.11%	30.64%	32.35%	31.13%	29.90%	30.15%	36.61%	35.88%	33.18%	38.24%	36.98%	35.48%
	Mel10	**22.83%**	23.91%	**22.83%**	20.83%	20.92%	**19.57%**	21.74%	20.11%	**18.48%**	24.73%	22.60%	**20.99%**

these ratios are generally expected to be close to each other.

In Table 8, when we examined the *Sc4 standing* activity case, where the lowest EER was 0.49%, the mean FAR was measured as 0.49% and the FRR was 0%. When we compare the results obtained from the activity-aware model with the EERs from Sect. 4.1, we can see that the EERs for all activities in the activity-aware model are generally lower than they were in Sect. 4.1. The number of training and validation images in Sect. 4.1 is higher than in the activity-aware model. Even in that case, the proposed framework shows better performance than direct biometric verification. Although Hexoskin and Faros devices achieved very close results in the activity-aware model, as in Sects. 4.1 and 4.2, it was observed that the Faros device was slightly more successful in general performance than the Hexoskin device in the ResNet50 model.

In Table 9, for the Faros device, the lowest EER results were obtained from the *walking* activity for all image types, unlike other models. However, as a general trend, the most successful activity is the *resting* activity. In addition, *1* sample cases (i.e. minimum genuine sample size in enrollment) for the same device have also achieved very successful results. As a general idea, using more samples yields lower EER results. However, the sample used in the recording is randomly selected and very successful results can be obtained in the *1* sample case since its discrimination might be higher among other images. For the Hexoskin device, it appears that the results are consistent with other devices and the Resnet50 model overall. The best performance on this device was obtained in Sc4 standing activity cases. It achieved 0.49% EER results for all genuine sample sizes (i.e. 1, 3 and 5). In addition, 0.49% FAR and 0% FRR were obtained in all cases. In this context, it has shown better performance than all other devices and models. When presenting both the best and worst outcomes, we observed the following results during the training of the DenseNet201 model, which we recommend for the Hexoskin device, using Mel2 images. The model achieved a validation accuracy rate of 97.40%. However, in the P11-P12 and 1 genuine sample case, the results were less favorable, with an EER of 46.47%, a FAR of 71.77%, and a FRR of 20%.

When comparing the Resnet50 and DenseNet201 models in terms of their results, it is typically observed that the DenseNet201 model yields lower EER results across all cases. Considering the best performances, the Faros device achieved 4.17% EER in the ResNet50 model and 4.29% EER in the DenseNet201 model in the Sc4 walking case, while the Hexoskin device achieved 0.49% EER in both models in the Sc4 standing case. From this observation, it can be concluded that the DenseNet201 model is more effective when used with the Hexoskin device while both CNN models obtained similar results for the Faros device.

Table 10 compares results from direct biometric verification and activity-aware models. To obtain Table 10, the mean of the EERs obtained for each activity in Tables 8 and 9 was calculated, and the EERs from Table 4 were subtracted from this value (i.e. *mean (Table* 8 *or Table* 9*) - Table* 4). When comparing Table 3 and Table 7, it becomes evident that the activity-aware biometric verification case utilied less training data. This is due to the fact that samples were divided based on their activities and analysed separately within the biometric verification model. Table 10 highlights the differences between the two cases and shows that the proposed biometric verification framework achieves better results.

5 Discussion

After examining ResNet50 and DenseNet201 CNN models, which are known to be highly effective DL studies, we tested various parameters. Activity-aware verification consistently achieved more successful results than direct biometric verification in both cases. Our study enables us to easily interpret the improvement in the performance of both medical and wearable devices.

The difference between activity-aware biometric verification and direct verification is most significant in spectrogram and scalogram images when using minimum sample cases (which are 2-s time windows and *1* genuine sample cases). However, the difference decreases as the number of genuine samples or the time window size increases. This indicates that the proposed biometric verification framework is suitable for real-life applications, particularly for short enrollment times. However, when analysing Mel-spectrogram images, it was found that the difference between activity-aware and direct biometric verification EERs is highest when using *5* genuine sample cases and 10-s time windows. This is due to the fact that Mel-spectrogram images contain less information compared to other image types. With an increase in time window size, more information can be obtained from the image, resulting in better results in activity-aware models during analysis.

It has been observed that EERs generally decrease as the enrollment time used in biometric verification increases. However, the best performances were obtained from *Sc4* cases. Our results demonstrate competitive outcomes in biometric verification in terms of EERs when compared to the results obtained from using the Deep-ECG [78], CNN+LSTM [90], EfficientNetB5 [91] and ECGXtractor [35] CNN models. In addition, these studies did not utilise short enrollment times. Obtaining low EERs even in short enrollment periods is vital for the real-life applicability of the proposed model. For example, if the proposed model is considered to be used on a smartwatch, the smartwatch can collect ECG data during the time it is worn and categorise this data according to activities. If ECG signals classified by activity are to be used for biometric verification (e.g. in the cases of waking the device from the sleep state or getting permission to access private data), it is important to investigate short enrollment times so that this process can be done quickly and with less error.

When we examine the activity classification, although the highest classification rate is achieved with the GoogleNet CNN model, when the mean performances are examined, DenseNet201 is seen as the most successful and GoogleNet as the most unsuccessful model. GoogleNet CNN model contains fewer learning layers than the other two models. The primary reasons for the lower accuracy rates in GoogLeNet are the limited amount of training data and the insufficient number of learning layers used for a classification task involving 4 activity classes.

When we examined the parameters in activity classification, 10 epochs yielded satisfactory results, while the 20 epochs cases were generally more successful. The SGDM optimiser produced higher results for the Hexoskin device, whereas the Adam optimiser achieved higher accuracy rates for the Faros device. In the DenseNet201 model, the highest accuracy rate was obtained from 2-s time windows, followed by 4-s time windows and 10-s time windows. However, in the ResNet50 and GoogleNet models, the order was 4 s, 2 s and 10 s time windows. For all the CNN models used, the highest measured accuracy rate was obtained from the *Sc4* case in the Faros device. Within the general biometric verification framework, we recommend using a 4-s time window and scalogram images for both devices. While both DenseNet201 and ResNet50 CNN models are suitable, it has been observed that the DenseNet201 model performs more successfully in wearable devices.

The images of the time-frequency representations show the heartbeat, which varies in quantities depending on the time window. For instance, in the 2-s time window cases, there is an average of 2 heartbeats, while in the 10-s time window cases, there is an average of 10 heartbeats. This shows us that more images are obtained in the 2-s case. Therefore, 10-s time windowed images have fewer training images, although they contain more heartbeat in each image. In the results, the lowest accuracy rates were generally obtained from images with 10-s time window cases. This shows that the number of images in the training set for activity classification is insufficient for the 10-s time window cases. In addition, the high accuracy rates of the 4-s time windows indicate that both the number of images

Table 10 Mean EER (%) differences between the direct biometric verification and the activity-aware verification

	ResNet50						**DenseNet201**					
Devices	**Faros**			**Hexoskin**			**Faros**			**Hexoskin**		
# of genuine samples	**1**	**3**	**5**	**1**	**3**	**5**	**1**	**3**	**5**	**1**	**3**	**5**
Image types												
Sc2	− 12.61%	− 8.79%	− 7.58%	− 15.55%	− 12.14%	− 11.64%	− 11.84%	− 7.81%	− 9.02%	− 12.95%	− 10.28%	− 11.27%
Sc4	− 7.93%	− 6.34%	− 4.56%	− 30.48%	− 9.10%	− 8.09%	− 15.72%	− 9.22%	−6.19%	− 19.67%	− 7.54%	− 5.70%
Sc10	− 6.51%	− 5.77%	− 5.39%	− 5.37%	− 6.03%	− 5.19%	− 6.46%	− 6.14%	−7.37%	− 7.42%	− 8.10%	− 7.12%
Sp2	− 7.90%	− 9.89%	− 8.89%	− 5.19%	− 1.79%	− 3.58%	− 7.72%	− 7.16%	− 7.35%	− 9.37%	− 8.20%	− 7.43%
Sp4	− 4.28%	− 2.02%	− 0.23%	− 12.22%	− 8.58%	− 7.88%	− 13.82%	− 9.48%	−9.49%	−12.03%	− 7.77%	− 7.49%
Sp10	− 7.64%	− 10.08%	− 10.89%	− 8.26%	− 7.34%	− 9.21%	−6.42%	−8.46%	− 9.00%	− 7.77%	− 8.45%	− 8.16%
Mel2	− 5.69%	− 3.38%	− 4.26%	− 0.16%	− 2.45%	− 3.19%	− 3.02%	− 2.56%	− 4.3%	− 1.93%	− 2.64%	− 4.18%
Mel4	− 8.14%	− 6.98%	− 5.27%	− 2.94%	− 1.51%	− 2.68%	− 6.00%	− 2.08%	− 2.27%	− 4.04%	− 3.86%	− 2.83%
Mel10	− 8.62%	− 10.68%	− 8.18%	− 2.94%	− 5.73%	− 4.88%	− 10.68%	− 12.61%	− 11.11%	− 6.64%	− 7.02%	−6.38%

used in training and the presence of more heartbeat (i.e. more information) in the images are important for the activity recognition model. Despite yielding lower results than our ML models [9] utilising the same dataset, our findings provide valuable insights into the comparative effectiveness of various time-frequency representations, time windows, epoch numbers, optimisers, and CNN models. Our study represents the first application of the Vollmer dataset [41] to activity classification.

For future research, training deep learning models from scratch for activity classification might improve accuracy. This is because using the ImageNet dataset with transfer learning did not yield successful results with time-frequency representations. Additionally, the created model can be tested using different devices with more activities and emotional states.

6 Conclusions

To conclude, the findings of this study suggest that the integration of ECG biometrics and activity classification based on ECG signals could potentially enhance authentication methods. The performance of wearable devices in this context was found to be comparable to that of medical devices. The development and application of this system in wearable technology like smartwatches could increase device authentication security and reduce the influence of daily activities on the authentication system.

Abbreviations

CNN	Convolutional neural networks
CWT	Continuous wavelet transformation
DAGNetwork	Directed acyclic graph network
DER	Detection error rate
DL	Deep learning
DWT	Discrete wavelet transform
ECG	Electrocardiogram
FAR	False Reject Rate
FAR	False accept rate
FC	Fully connected
Inf-FS	Infinite feature selection
MFCC	Mel frequency cepstrum coefficient
ML	Machine learning
SGDM	Stochastic gradient descent with momentum
STFT	Short-time Fourier transform
SWT	Stationary wavelet transform

Acknowledgements

Hazal Su Bıçakcı Yeşilkaya would like to thank the Ministry of National Education of the Republic of Turkiye for the PhD fee being covered. However, since this study uses a public dataset and the project is carried out within the University of Kent, this project does not have a sponsor.

Authors' contributions

H.S.B.Y. wrote the main manuscript text, created all figures and analyzed all experiments. R.G. contributed to the study by developing the idea for the project, executing the project and acting as a supervisor. All authors reviewed and edited the manuscript.

Data availability

The data used is open access. Source is on the following site: https://physionet.org/content/simultaneous-measurements/1.0.0/.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

The authors permit this work to be published in the EURASIP Journal on Information Security, and other publications produced by the EURASIP Journal on Information Security, in print and online.

Competing interests

The authors declare that they have no competing interests.

Received: 25 July 2024 Accepted: 12 February 2025
Published online: 25 February 2025

References

1. S.A. Israel, J.M. Irvine, A. Cheng, M.D. Wiederhold, B.K. Wiederhold, Ecg to identify individuals. Pattern Recog. **38**(1), 133–142 (2005)
2. S. Yin, M. Kim, D. Kadetotad, Y. Liu, C. Bae, S.J. Kim, Y. Cao, J.s. Seo, A 1.06-μw smart ecg processor in 65-nm cmos for real-time biometric authentication and personal cardiac monitoring. IEEE J. Solid-State Circ. **54**(8), 2316–2326 (2019)
3. ISO iso/iec 2382-37:2022 information technology - vocabulary - part 37: Biometrics. https://www.iso.org/standard/73514.html. Accessed 16 Jan 2025
4. S. Migas, M.L. Ellis, B. Wrona, E. Rivero Sanz, J. Brownrigg, O. Strauss, F.Z. Ahmed, Missed opportunities in heart failure diagnosis and management: study of an urban UK population. ESC Heart Fail. **11**(4), 2200–2213 (2024)
5. P. Kamga, R. Mostafa, S. Zafar, The use of wearable ecg devices in the clinical setting: a review. Curr. Emerg. Hosp. Med. Rep. **10**(3), 67–72 (2022)
6. K. Hibbitt, J. Brimicombe, M.R. Cowie, A. Dymond, B. Freedman, S.J. Griffin, F.D.R.J. Hobbs, H.C. Lindén, G.Y.H. Lip, J. Mant, R.J. McManus, M. Pandiaraja, K. Williams, P.H. Charlton, Reliability of single-lead electrocardiogram interpretation to detect atrial fibrillation: insights from the SAFER feasibility study. EP Europace **26**(7), euae181 (2024). https://doi.org/10.1093/europace/euae181
7. A.S. Jat, T.M. Grønli, in *International Work-Conference on Bioinformatics and Biomedical Engineering*. Smart watch for smart health monitoring: a literature review (Springer International Publishing, Cham, 2022), pp. 256–268
8. D. Hernando, N. Garatachea, R. Almeida, J.A. Casajús, R. Bailón, Validation of heart rate monitor polar rs800 for heart rate variability analysis during exercise. J Strength Cond. Res. **32**(3), 716–725 (2018)
9. H.S. Bıçakcı, M. Santopietro, R. Guest, Activity-based electrocardiogram biometric verification using wearable devices. IET Biom. **12**(1), 38–51 (2023). https://doi.org/10.1049/bme2.12105
10. H.S. Bıçakcı, M. Santopietro, M. Boakes, R. Guest, in *2021 International Carnahan Conference on Security Technology (ICCST)*. Evaluation of electrocardiogram biometric verification models based on short enrollment time on medical and wearable recorders (2021), pp. 1–6. https://doi.org/10.1109/ICCST49569.2021.9717372
11. E.S. Dahiya, A.M. Kalra, A. Lowe, G. Anand, Wearable technology for monitoring electrocardiograms (ecgs) in adults: a scoping review. Sensors **24**(4), 1318 (2024)
12. R. Fonkou, R. Kengne, M. Wamba, H.C.F. Kamgang, P. Talla, On the heart rhythm analysis using a nonlinear dynamics perspective: analytical study and electronic simulation. Phys. Scr. **99**(5), 055270 (2024)
13. Y. Li, Y. Pang, K. Wang, X. Li, Toward improving ECG biometric identification using cascaded convolutional neural networks. Neurocomputing **391**, 83–95 (2020)
14. D.A. AlDuwaile, M.S. Islam, Using convolutional neural network and a single heartbeat for ecg biometric recognition. Entropy **23**(6), 733 (2021)

15. R. Begum, A. Sharma, G. Singh, Ecg based reliable user identification using deep learning. Int. J. Biomed. Biol. Eng. **16**(9), 66–75 (2022)
16. Y.H. Byeon, S.B. Pan, K.C. Kwak, Intelligent deep models based on scalograms of electrocardiogram signals for biometrics. Sensors **19**(4), 935 (2019)
17. Physionet ptb diagnostic ecg database. https://physionet.org/content/ptbdb/1.0.0/. Accessed 18 Jan 2025
18. Physionetcu ventricular tachyarrhythmia database. https://physionet.org/content/cudb/1.0.0/. Accessed 18 Jan 2025
19. I.B. Ciocoiu, N. Cleju, Off-person ecg biometrics using spatial representations and convolutional neural networks. IEEE Access **8**, 218966–218981 (2020). https://doi.org/10.1109/ACCESS.2020.3042547
20. S. Pouryayevali, S. Wahabi, S. Hari, D. Hatzinakos, in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. On establishing evaluation standards for ecg biometrics (IEEE, 2014), pp. 3774–3778
21. H.P. da Silva, A. Lourenço, A. Fred, N. Raposo, M.A. de Sousa, Check your biosignals here: A new dataset for off-the-person ecg biometrics. Comput. Methods Programs Biomed. **113**(2), 503–514 (2014). https://doi.org/10.1016/j.cmpb.2013.11.017
22. Physionetfantasia database. https://physionet.org/content/fantasia/1.0.0/. [Online], Accessed 18 Jan 2025
23. Physionetcombined measurement of ecg, breathing and seismocardiograms. https://physionet.org/content/cebsdb/1.0.0/. Accessed 18 Jan 2025
24. Physionetmit-bih normal sinus rhythm database. https://physionet.org/content/nsrdb/1.0.0/. Accessed 18 Jan 2025
25. Physionetmit-bih st change database. https://physionet.org/content/stdb/1.0.0/. Accessed 18 Jan 2025
26. Physionetmit-bih atrial fibrillation database. https://physionet.org/content/afdb/1.0.0/. Accessed 18 Jan 2025
27. Physionetecg-id database. https://physionet.org/content/ecgiddb/1.0.0/. Accessed 18 Jan 2025
28. M. Hammad, P. Pławiak, K. Wang, U.R. Acharya, Resnet-attention model for human authentication using ecg signals. Expert Syst. **38**(6), e12547 (2021)
29. Physionetmit-bih arrhythmia. https://physionet.org/content/mitdb/1.0.0/. Accessed 18 Jan 2025
30. Physionetqt database. https://physionet.org/content/qtdb/1.0.0/. Accessed 18 Jan 2025
31. Physionetintracardiac atrial fibrillation database. https://physionet.org/content/iafdb/1.0.0/. Accessed 18 Jan 2025
32. Physionetmimic ii databases. https://archive.physionet.org/mimic2/. Accessed 18 Jan 2025
33. Physionetmimic-iii waveform database. https://physionet.org/content/mimic3wdb/1.0/. Accessed 18 Jan 2025
34. M. Hammad, S. Zhang, K. Wang, A novel two-dimensional ecg feature extraction and classification algorithm based on convolution neural network for human authentication. Futur. Gener. Comput. Syst. **101**, 180–196 (2019). https://doi.org/10.1016/j.future.2019.06.008
35. P. Melzi, R. Tolosana, R. Vera-Rodriguez, Ecg biometric recognition: Review, system proposal, and benchmark evaluation. IEEE Access (2023)
36. A. Sanz-García, A. Cecconi, A. Vera, J.M. Camarasaltas, F. Alfonso, G.J. Ortega, J. Jimenez-Borreguero, Electrocardiographic biomarkers to predict atrial fibrillation in sinus rhythm electrocardiograms. Heart **107**(22), 1813–1819 (2021)
37. P. Melzi, R. Tolosana, A. Cecconi, A. Sanz-Garcia, G.J. Ortega, L.J. Jimenez-Borreguero, R. Vera-Rodriguez, Analyzing artificial intelligence systems for the prediction of atrial fibrillation from sinus-rhythm ecgs including demographics and feature visualization. Sci. Rep. **11**(1), 22786 (2021)
38. A.J. Prakash, K.K. Patro, S. Samantray, P. Pławiak, M. Hammad, A deep learning technique for biometric authentication using ecg beat template matching. Information **14**(2), 65 (2023)
39. B. Abd El-Rahiem, M. Hammad, A multi-fusion iot authentication system based on internal deep fusion of ecg signals. Security and Privacy Preserving for IoT and 5G Networks: Techniques, Challenges, and New Directions. (Springer Nature Switzerland, 2022). pp. 53–79
40. Y.H. Byeon, K.C. Kwak, Pre-configured deep convolutional neural networks with various time-frequency representations for biometrics from ecg signals. Appl. Sci. **9**(22), 4810 (2019)
41. M. Vollmer, D. Bläsing, L. Kaderali, in *2019 Computing in Cardiology (CinC)*. Alignment of multi-sensored data: Adjustment of sampling frequencies and time shifts (IEEE, Singapore, 2019), p. 1
42. S. Wahabi, S. Pouryayevali, S. Hari, D. Hatzinakos, On evaluating ecg biometric systems: Session-dependence and body posture. IEEE Trans. Inf. Forensic Secur. **9**(11), 2002–2013 (2014). https://doi.org/10.1109/TIFS.2014.2360430
43. S. Wahabi, S. Pouryayevali, D. Hatzinakos, in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Posture-invariant ecg recognition with posture detection (2015), pp. 1812–1816. https://doi.org/10.1109/ICASSP.2015.7178283
44. J. Kim, D. Sung, M. Koh, J. Kim, K.S. Park, Electrocardiogram authentication method robust to dynamic morphological conditions. IET Biom. **8**(6), 401–410 (2019)
45. J. Liu, J. Chen, H. Jiang, W. Jia, Q. Lin, Z. Wang, in *2018 IEEE international symposium on circuits and systems (ISCAS)*. Activity recognition in wearable ecg monitoring aided by accelerometer data (IEEE, Florence, 2018), pp. 1–4
46. F.S. Butt, L. La Blunda, M.F. Wagner, J. Schäfer, I. Medina-Bulo, D. Gómez-Ullate, Fall detection from electrocardiogram (ecg) signals and classification by deep transfer learning. Information **12**(2), (2021). https://doi.org/10.3390/info12020063
47. G. Cosoli, L. Antognoli, L. Scalise, Wearable electrocardiography for physical activity monitoring: Definition of validation protocol and automatic classification. Biosensors **13**(2), 154 (2023)
48. M.M.M. Nawawi, K.A. Sidek, A.W. Azman, Ecg biometric in real-life settings: analysing different physiological conditions with wearable smart textiles shirts. Bull. Electr. Eng. Inform. **12**(5), 2930–2938 (2023)
49. D. Kim, B. Heo, D. Han, in *European Conference on Computer Vision*. Densenets reloaded: paradigm shift beyond resnets and vits (Springer Nature Switzerland, Cham, 2024), pp. 395–415
50. C. Zhang, P. Benz, D.M. Argaw, S. Lee, J. Kim, F. Rameau, J.C. Bazin, I.S. Kweon, in *Proceedings of the IEEE/CVF winter conference on applications of computer vision (WACV), Waikoloa, HI, USA*. Resnet or densenet? Introducing dense shortcuts to resnet (IEEE, 2021), pp. 3549–3558. https://doi.org/10.1109/WACV48630.2021.00359
51. M. Taki, Deep residual networks and weight initialization. ArXiv (2017). arXiv preprint arXiv:1709.02956
52. G. Huang, Z. Liu, G. Pleiss, L. Van Der Maaten, K.Q. Weinberger, Convolutional networks with dense connectivity. IEEE Trans. Pattern Anal. Mach. Intell. **44**(12), 8704–8716 (2019)
53. L. Alzubaidi, J. Zhang, A.J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M.A. Fadhel, M. Al-Amidie, L. Farhan, Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. J. Big Data **8**, 1–74 (2021)
54. C.T. Lu, C.J. Ou, Y.Y. Lu, A practical app for quickly calculating the number of people using machine learning and convolutional neural networks. Appl. Sci. **12**(12), 6239 (2022)
55. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, in *Proceedings of the IEEE conference on computer vision and pattern recognition*. Going deeper with convolutions (IEEE, Boston, 2015), pp. 1–9
56. D.P. Kingma, J. Ba, Adam: A method for stochastic optimization. 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA (2014) arXiv preprint arXiv:1412.6980
57. S.R. Dubey, S. Chakraborty, S.K. Roy, S. Mukherjee, S.K. Singh, B.B. Chaudhuri, diffgrad: An optimization method for convolutional neural networks. IEEE Trans. Neural Netw. Learn. Syst. **31**(11), 4500–4511 (2020). https://doi.org/10.1109/TNNLS.2019.2955777
58. Y. Chen, H. Ge, X. Su, X. Ma, Classification of exercise fatigue levels by multi-class svm from ecg and hrv. Med. Biol. Eng. Comput. **62**, 2853–2865 (2024)
59. M. Abdel-Latif, M.R. Askari, M.M. Rashid, M. Park, L. Sharp, L. Quinn, A. Cinar, Multi-task classification of physical activity and acute psychological stress for advanced diabetes treatment. Signals **4**(1), 167–192 (2023). https://doi.org/10.3390/signals4010009
60. S. Thompson, P. Fergus, C. Chalmers, D. Reilly, in *2020 International Joint Conference on Neural Networks (IJCNN)*. Detection of obstructive sleep apnoea using features extracted from segmented time-series ecg signals using a one dimensional convolutional neural network (2020), pp. 1–8. https://doi.org/10.1109/IJCNN48605.2020.9207470

61. M. Vollmer, D. Bl"asing, J.E. Reiser, M. Nisser, A. Buder, Simultaneous physiological measurements with five devices at different cognitive and physical loads, version 1.0.0 ed. Physionet **2020**, (2020). https://doi.org/10.13026/chd5-t946
62. A.L. Goldberger, L.A. Amaral, L. Glass, J.M. Hausdorff, P.C. Ivanov, R.G. Mark, J.E. Mietus, G.B. Moody, C.K. Peng, H.E. Stanley, Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. Circulation **101**(23), e215–e220 (2000)
63. Mind Media Neuro and Biofeedback Systems nexus-10 mkii. https://www.mindmedia.com/en/products/nexus-10-mkii/. (2016). Accessed 01 Feb 2023
64. Bittium Corporation bittium emotion faros 360°. https://shop.bittium.com/product/37/emotion-faros-360. (2017). Accessed 01 Feb 2023
65. G. Bilo, C. Zorzi, J.E.O. Munera, C. Torlasco, V. Giuli, G. Parati, Validation of the somnotouch-nibp noninvasive continuous blood pressure monitor according to the european society of hypertension international protocol revision 2010. Blood Press. Monit. **20**(5), 291 (2015)
66. H.H. Sensors, AI, Hexoskin health sensors & ai (2015). https://www.hexoskin.com/pages/health-research. [Online] Accessed 01 Feb 2023
67. Y. Li, G. Liu, in *2016 IEEE International Conference on Network Infrastructure and Digital Content (IC-NIDC)*. Sound classification based on spectrogram for surveillance applications (2016), pp. 293–297. https://doi.org/10.1109/ICNIDC.2016.7974583
68. B. Boashash, *Time-frequency signal analysis and processing: a comprehensive reference*. Academic press in Elsevier (2015)
69. H. Li, P. Boulanger, Structural anomalies detection from electrocardiogram (ecg) with spectrogram and handcrafted features. Sensors **22**(7), (2022). https://doi.org/10.3390/s22072467
70. Scalogram Computation in Signal Analyzer mathworks. https://uk.mathworks.com/help/signal/ug/scalogram-computation-in-signal-analyzer.html. (2023). Accessed 02 Mar 2023
71. H. Khorrami, M. Moavenian, A comparative study of dwt, cwt and dct transformations in ecg arrhythmias classification. Expert Syst. Appl. **37**(8), 5751–5757 (2010)
72. Continuous wavelet transform filter bank mathworks. https://uk.mathworks.com/help/wavelet/ref/cwtfilterbank.html. (2023). Accessed 02 Mar 2023
73. W.J. Poser, Douglas O'Shaughnessy, speech communication: Human and machine. reading, massachusetts: Addison-wesley publishing company, 1987. pp. xviii 568. isbn 0-201-16520-1. J. Int. Phon. Assoc. **20**(2), 52–54 (1990). https://doi.org/10.1017/S002510030000431X
74. H. Meng, T. Yan, F. Yuan, H. Wei, Speech emotion recognition from 3d log-mel spectrograms with deep learning network. IEEE Access **7**, 125868–125881 (2019). https://doi.org/10.1109/ACCESS.2019.2938007
75. Mel Spectrogram mathworks. https://uk.mathworks.com/help/audio/ref/melspectrogram.html. (2023). Accessed 07 May 2023
76. A.A. Aleidan, Q. Abbas, Y. Daadaa, I. Qureshi, G. Perumal, M.E. Ibrahim, A.E. Ahmed, Biometric-based human identification using ensemble-based technique and ecg signals. Appl. Sci. **13**(16), 9454 (2023)
77. J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, L. Fei-Fei, in *2009 IEEE conference on computer vision and pattern recognition*. Imagenet: A large-scale hierarchical image database. (IEEE, Miami, FL, 2009), pp. 248–255
78. R. Donida Labati, E. Muñoz, V. Piuri, R. Sassi, F. Scotti, Deep-ecg: Convolutional neural networks for ecg biometric recognition. Pattern Recogn. Lett. **126**, 78–85 (2019). https://doi.org/10.1016/j.patrec.2018.03.028. Robustness, Security and Regulation Aspects in Current Biometric Systems
79. Y. Sun, Y. Chen, X. Wang, X. Tang, in *Proceedings of the 27th International Conference on Neural Information Processing Systems*. Deep learning face representation by joint identification-verification, vol. 2, (2014), pp. 1988–1996
80. Matlab MathWorks options for training deep learning neural network. (2023). https://uk.mathworks.com/help/deeplearning/ref/trainingoptions.html. Accessed 29 May 2023
81. S. Mittal, P. Rajput, S. Subramoney, A survey of deep learning on cpus: opportunities and co-optimizations. IEEE Trans Neural Netw. Learn. Syst. **33**(10), 5095–5115 (2021)
82. K. He, X. Zhang, S. Ren, J. Sun, in *Proceedings of the IEEE conference on computer vision and pattern recognition*. Deep residual learning for image recognition, (IEEE, Las Vegas, NV, 2016), pp. 770–778
83. R. Gomes, C. Kamrowski, J. Langlois, P. Rozario, I. Dircks, K. Grottodden, M. Martinez, W.Z. Tee, K. Sargeant, C. LaFleur et al., A comprehensive review of machine learning used to combat covid-19. Diagnostics **12**(8), 1853 (2022)
84. J. Allgaier, R. Pryss, Cross-validation visualized: A narrative guide to advanced methods. Mach. Learn. Knowl. Extraction **6**(2), 1378–1388 (2024). https://doi.org/10.3390/make6020065
85. MatLab MathWorks directed acyclic graph (dag) network for deep learning. https://uk.mathworks.com/help/deeplearning/ref/dagnetwork.html. Accessed 30 May 2023
86. MatLab MathWorks what is the difference between layergraph and dagnetwork in deep learning? https://uk.mathworks.com/matlabcentral/answers/409833-what-is-the-difference-between-layergraph-and-dagnetwork-in-deep-learning. Accessed 30 May 2023
87. M. Santopietro, An exploration of dynamic biometric performance using device interaction and wearable technologies. Ph.D. thesis, University of Kent, (2022). https://kar.kent.ac.uk/98627/. Accessed 18 Jan 2025
88. S.Y. ŞEN, N. ÖZKURT, in *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*. Convolutional neural network hyperparameter tuning with adam optimizer for ecg classification (IEEE, Istanbul, 2020), pp. 1–6
89. M. Mahsereci, L. Balles, C. Lassner, P. Hennig. Early stopping without a validation set. (2017). https://arxiv.org/abs/1703.09580 . Accessed 18 Jan 2025
90. P. Tirado-Martin, R. Sanchez-Reillo, Bioecg: Improving ecg biometrics with deep learning and enhanced datasets. Appl. Sci. **11**(13), 5880 (2021)
91. N. Ammour, R.M. Jomaa, M.S. Islam, Y. Bazi, H. Alhichri, N. Alajlan, Deep contrastive learning-based model for ecg biometrics. Appl. Sci. **13**(5), 3070 (2023)

Publisher's Note