A novel framework for ECG biometric verification on mobile devices utilising activity classification

Hazal Su Bıçakcı

A Thesis submitted to the University of Kent for the Degree of Doctor of Philosophy in Electronic Engineering

School of Engineering
University of Kent

Pages: 195
October 2023.

Abstract

Considering the digital world, almost all online transactions (e.g. payments, shopping, etc.) need identity verification and information security. In addition to models such as traditional passwords and PINs, biometrics such as facial recognition and fingerprints have also begun to be used frequently for this purpose. The literature has often stated that electrocardiogram (ECG) biometrics can also provide reliable results and increase performance in multi-models. However, it has also been stated that the characteristics of ECG signals change over time due to environmental, biological or physiological reasons such as physical activities and emotional states. This affects the performance and stability of the model.

Another open challenge is the need for difficult-to-use devices and sensors to collect ECG data. With the development of wearable devices, many studies have evaluated the performance of these devices. However, to provide a reliable suitable real-life scenarios ECG-based biometric verification model, many parameters must be investigated in depth.

The scope of this thesis is to examine the parameters affecting ECG biometrics in verification models, to create a novel framework to increase long-term stability and to test the created framework on various mobile devices.

The contribution of this thesis to the literature is by creating an activity-aware and emotional status-aware biometric verification framework, demonstrating the usability of this framework for medical and wearable devices.

Acknowledgements

I wish to convey my sincere gratitude to Professor Richard Guest, my supervisor, for his constant support, patience, motivation, and extensive knowledge. His guidance was priceless in every step of the research and writing journey for my thesis.

My gratitude also extends to my review panel chair, Dr. Sanaul Hoque, for his support and insightful feedback, which greatly enhanced the quality and lucidity of my work.

I am profoundly indebted to my family, Hatice and İlhan Bıçakcı, my better half Bartu Yeşilkaya, as well as my friends Fatma Burcu Bozan, Jale Kandemir, Aytülü Akbaşlı, Nuray Demir, Ezgi Yalçıner, Başak Şimşek and Kathryn Heffner for their constant support and assistance during challenging times.

I have had the privilege of working with a supportive group of PhD researchers. I extend my thanks to Dr. Marco Santopietro, Dr. Matthew Boakes and all my colleagues for making this journey enjoyable.

Last but not least, I would like to acknowledge the Republic of Turkiye Ministry of National Education for their sponsorship. I am deeply grateful to Mustafa Kemal Atatürk, the founder of the Republic of Turkiye who started this scholarship on 16/04/1929 and spent his whole life on the development of Turkiye.

Hazal Su Bıçakcı, August 2024.

Published Content

The main outcome of this thesis is to propose a new framework for ECG biometrics. The following two publications were made in a peer-reviewed journal and at a peer-reviewed conference during this research:

- "*Evaluation of Electrocardiogram Biometric Verification Models Based on Short Enrollment Time on Medical and Wearable Recorders*", Hazal Su Bıçakcı, Marco Santopietro, Matthew Boakes and Richard Guest, *2021 International Carnahan Conference on Security Technology (ICCST)*, 2021, pp. 1-6.

This publication [1] explores the effects of different enrollment times on biometric verification. The performance of medical and wearable devices in biometric verification has been compared using many subjects and has been the basis for our studies. The methodology used is described in Chapter 3 and the results are described in Chapter 4.

- "*Activity-based Electrocardiogram Biometric Verification Using Wearable Devices*", Hazal Su Bıçakcı, Marco Santopietro and Richard Guest, *IET Biometrics*, Volume 12, Issue 1, pp. 38-51, 2023.

In this paper [2], we present our study focused on the activity-aware biometric verification model we have developed, along with the novel features we have incorporated. The methodology used is explained in Chapter 3, direct verification results in Chapter 4, activity and emotional status classification methods and results in Chapter 5 and activity-aware biometric verification results in Chapter 6.

Our third publication on the use of the proposed framework in the deep learning model is under evaluation. Parts of the above publications appear verbatim in this thesis.

List of Figures

List of Tables

List of Acronyms

AC Auto-Correlation

Adam Adaptive Moment Estimation

AI Artificial intelligence

AV Atrioventricular node

BN Batch Normalization

Bpm Beat-per-minute

BT Bagged Tree Ensemble

CNN Convolutional Neural Networks

CWT Continuous Wavelet Transformation

DAGNetwork Directed Acyclic Graph Network

DCT Discrete Cosine Transformation

DER Detection Error Rate

DL Deep Learning

DNN Deep Neural Networks

DT Decision Trees

DTW Dynamic Time Wrapping

DWT Discrete Wavelet Transform

E-HOL E-HOL-03-0202-003

ECG Electrocardiogram

EDA Electrodermal activity

EER Equal Error Rate

EMG Electromyography

FAR False Accept Rates

FC Fully Connected

FIR Finite Impulse Response

FN False Negative

FP False Positive

FRR False Reject Rates

FT Fine-tuned

GRU Gated Recurrent Unit

GSR Galvanic Skin Response

HR Heart Rate

IEC International Electrotechnical Commission

IIR Infinite Impulse Response

ISO International Organization for Standardisation

KNN k-Nearest Neighbours

LDA Linear Discriminant Analysis

LOSO Leave-one-subject-out

LSTM Long Short-Term Memory

MCC MultiClass Classifier

MFCC Mel frequency cepstrum coefficient

ML Machine Learning

MLP Multi-Layer Perceptron

MMSF Multimodal-Multisensory Sequential Fusion

NB Naive Bayes

NN Neural Networks

RFC Random Forest

S-KNN Subspace KNN Ensemble

SA Sinoatrial node

Sgdm Stochastic Gradient Descent with Momentum

SSD Sum of Squared Difference

SVM Support Vector Machines

SWT Stationary Wavelet Transform

TN True Negative

TP True Positive

Vollmer Simultaneous physiological measurements with five devices at different cognitive and physical loads

WeSAD Multimodal Dataset for Wearable Stress and Affect Detection

Contents

Chapter 1

Introduction

The history of biometrics has encompassed the use of measurable characteristics for security purposes and forensic cases such as manual face recognition, comparing body measurements and fingerprint comparison. Moreover, fingerprints were used to sign contracts in history because fingerprints were a way for individuals to prove their identity. With the rapid development of mobile devices and technology, the area of usage for biometrics has expanded and has been generally accepted worldwide. Today, there are three basic methods for people to prove their identity. These are something they have, such as cards (ID, passports and bank cards), specific known information they have such as passwords, or biometric attributes such as fingerprints.

In today's world of rapid technological advancement and heightened concerns for information security and privacy, biometric authentication has become a vital field of significant importance for various industries such as online banking and airport border controls. Biometrics, which involves analysing an individual's unique physical and behavioural traits, has received substantial attention due to its potential to transform identification and authentication processes.

According to the International Organization for Standardisation (ISO) and International Electrotechnical Commission (IEC), there are two main categories of biometrics: physiological and behavioural [6]. Physiological biometrics refer to

unique features of the human body, such as fingerprints, facial structure and iris. Behavioural biometrics refer to acquired knowledge and behaviours, such as gait, hand gestures and keystroke dynamics. Some types of biometrics can show both physiological and behavioural characteristics (shown as a *"Hybrid"* in Fig.1.1). For example, an Electrocardiogram (ECG) is a physiological signal by nature, but it is included in both biometric groups as it is affected by many behavioural characteristics such as stress, emotional status and physical activities. The types of biometrics for identity confirmation of individuals are shown in Fig.1.1.

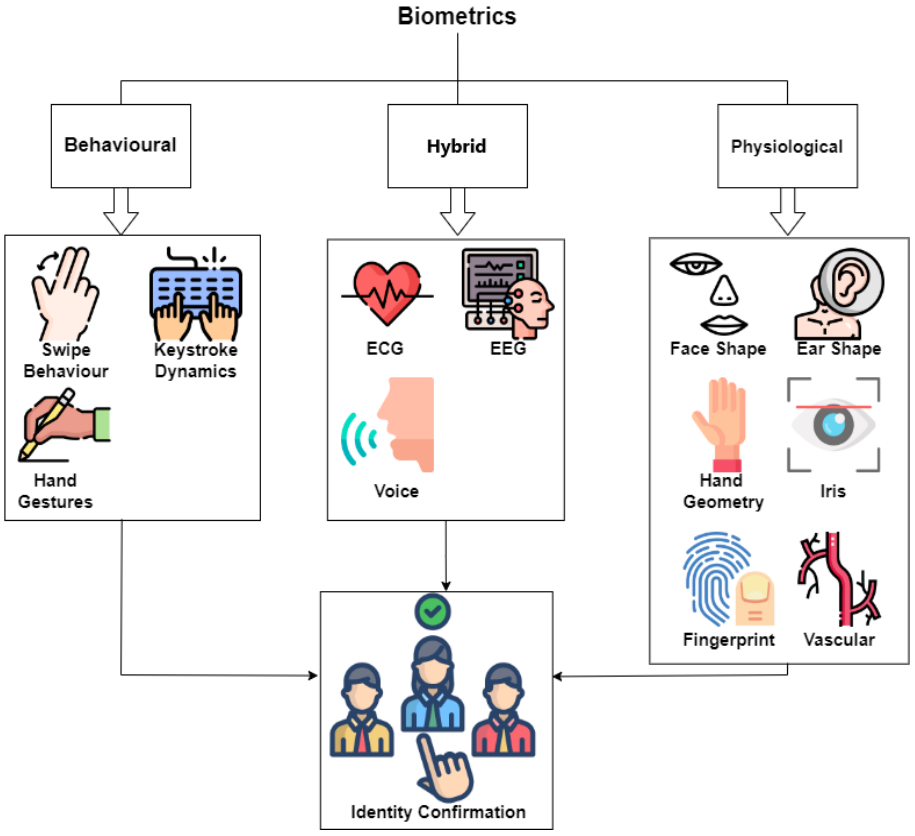

FIGURE 1.1: Identity confirmation of individuals

There are several advantages and disadvantages associated with both types of biometrics. For instance, biometric features are inherently linked to an individual and cannot be separated, lost, or transferred to another person. In addition, biometric recognition or verification models generally exhibit high success rates [7]. However, the collection of biometric data requires specialised environments, sensors, and devices. Furthermore, the data collected may vary depending on the

environment, condition, health status, etc., which might require the collection of extensive personal data.

There are two ways to use biometric data to confirm a person's identity. One of them is biometric verification, a 1-to-1 matching of claimed identity and that identity's enrolled record in a database. Biometric verification can be used for access controls such as accessing computers, mobile phones and specific buildings. Biometric authentication refers to the continuous validation process. While verification is usually performed once to confirm the identity initially, authentication is an ongoing process that continually validates that identity. For instance, within a high-security building, facial recognition cameras might authenticate a user's identity multiple times as they navigate through various areas. The other method is biometric identification which is a 1-to-N matching of one person's biometric data to all subjects' biometric data in a database. Biometric identification can be used for law enforcement such as identifying suspects or criminals using CCTV cameras.

In this work, the access performance of mobile devices was investigated using ECG data, and since this process is a one-time (non-ongoing) process, biometric verification was used. General outlines of ECG-based biometric verification and identification are illustrated in Fig.1.2.

Not all features of the human body are considered biometrics. There are specific criteria that signals or images must meet to be classified as biometrics. Jain et al. [8, 9] defined these criteria as universality, uniqueness/distinctiveness, permanence, collectability/measurability, performance, acceptability and circumvention. The ECG biometric model, which includes these parameters, can be described in the following way:

- **Universality:** The ECG signal is observable in all living creatures so it is universal.

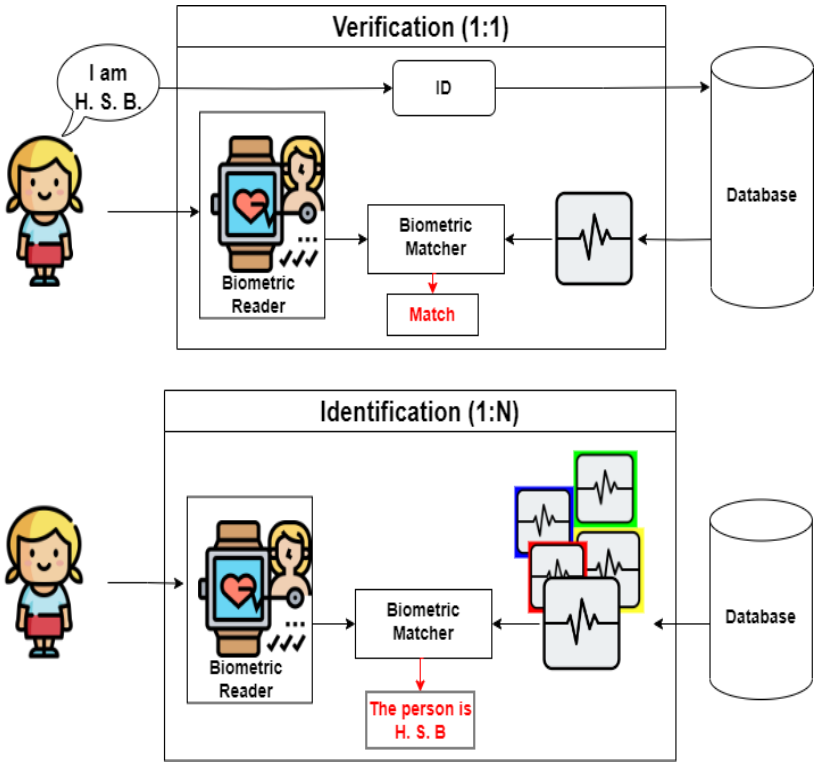

FIGURE 1.2: Outlines of biometric applications

- **Uniqueness/Distinctiveness:** Each individual has a unique ECG wave pattern that can be identified by the slight variations in peak and trough amplitudes and distances between them [10].

- **Permanence:** ECG signals exhibit a stable structure that remains consistent for an extended duration (especially for healthy subjects), meeting this criterion. However, although the general heartbeat pattern is the same, unstable measurements may also be seen in the long term and according to environmental factors.

- **Collectability/Measurability:** Wearable devices and sensors now make it easy to measure ECG thanks to advancements in wearable technology.

- **Performance:** Biometric verification and identification models based on ECG can achieve impressive results [11]. Although each biometric model has pros and cons, ECG biometrics is reliable in the general performance evaluation.

- **Acceptability:** Thanks to wearable devices such as smartwatches, it is easy to obtain ECG data. People are often willing to share their ECG data for purposes such as monitoring their own health.

- **Circumvention:** The use of a highly secure authentication solution is crucial to prevent circumvention. ECG signals are hardly imitable so it is an ideal solution to build secure biometric models. Thus, ECG is generally used in multimodel biometrics to increase security [12, 13]. However, the ECG-based model may be weak in replay attacks. The solution to this is encryption methods.

In Section 1.1, we will explain ECG signals and how the ECG biometric model compares to other models.

1.1 ECG Biometrics

By measuring the electrical activity of the heart through electrodes attached to the human body, an ECG can identify the unique heartbeat rhythm of humans. This feature makes ECG an ideal candidate for authentication systems [4, 14–16]. The performance comparison of the ECG biometrics with other biometric models is shown in Table 1.1 [17]. The comparisons in this table change over time with advancing device and sensor technology and the variety of methods used. Thus, Koffi et al. [17] updated this table for 15 biometric traits in 2023. However, since there was no evaluation for ECG biometrics, an addition was made by us.

Wearable and medical devices offer various options for collecting ECG recordings, including different electrode types (dry or wet) and a range of electrode numbers (1, 2, 3, 6, or 12 leads). The placement of the electrodes is also crucial for accurate readings [2]. There are many devices available, including chest bands [18, 19], armbands [20], wristbands [21], smartwatches [22] and Holter monitors [23], that can measure ECG from different areas of the body. In a classical ECG recording, there are 12 leads attached to the human body. However, there is only one sensor

TABLE 1.1: Performance comparison of different biometric models

Biometric Traits	Universality	Uniqueness	Permanence	Collectability	Performance	Acceptability	Circumvention
DNA	High	High	High	Low	High	Low	Low
Ear	Medium	Medium	High	Medium	Medium	High	Medium
Face	High	Low	Medium	High	Low	High	High
Facial Thermal	High	High	Low	High	Medium	High	Low
Fingerprint	Medium	High	High	Medium	High	Medium	Medium
Gait	Medium	Low	Low	High	Low	High	Medium
Hand Geometry	Medium	Medium	Medium	High	Medium	Medium	Medium
Hand Vein	Medium	Medium	Medium	Medium	Medium	Medium	Low
Iris	Medium	High	High	Medium	High	Low	Low
Keystroke	Low	Low	Low	Medium	Low	Medium	Medium
Odor	High	High	High	Low	Low	Medium	Low
Palmprint	Medium	High	High	Medium	High	Medium	Medium
Retina	High	High	Medium	Low	High	Low	Low
Signature	Low	Low	Low	High	Low	High	High
Voice	High	High	Medium	Medium	Medium	High	Low
ECG	High	High	Medium	High	Medium	Medium	Low

on wearable devices. The number of electrodes, sensors and their locations affect the usability of the system.

1.1.1 Overview of ECG

The heart basically consists of four parts. The upper chamber of the heart is called the atrial and the lower chamber is called the ventricular, and there is one atrial and one ventricular on the right and left sides of the heart. The regular beating of the heart is controlled by two nerves in the heart, which are the Sinoatrial node (SA) and the Atrioventricular node (AV). These nodes work as pacemakers of the heart. The right atrium's upper wall contains the SA node. Both atria contract as a result of nerve impulses that are produced and pass across the heart wall [24]. Near the bottom of the right atrium, on the right side of the wall separating the atria, is the AV node. The AV node receives the impulses produced by the SA node after a tenth of a second delay. The atria contract during this interval, transferring the blood into the ventricles prior to ventricular contraction

(i.e. forming the P peak). The atrioventricular bundle (i.e. His bundle) then delivers impulses to the Purkinje fibres in ventricles via the AV node (i.e. forming Q, R and S peaks). The AV node controls electrical signals to prevent electrical impulses from moving too quickly, which can cause atrial fibrillation [24]. In the end, ventricular relaxation and repolarisation occur, forming the T peak. This cardiac cycle repeats with regularity unless there is another effect such as physical activity, medical drug interventions and heart failures. To better comprehend heart conditions and ECG-based applications, various essential characteristics are extracted from the ECG waveform. The P, Q, R, S and T peaks constitute each heartbeat. A single heartbeat representation is shown in Fig.1.3 [25, 26].

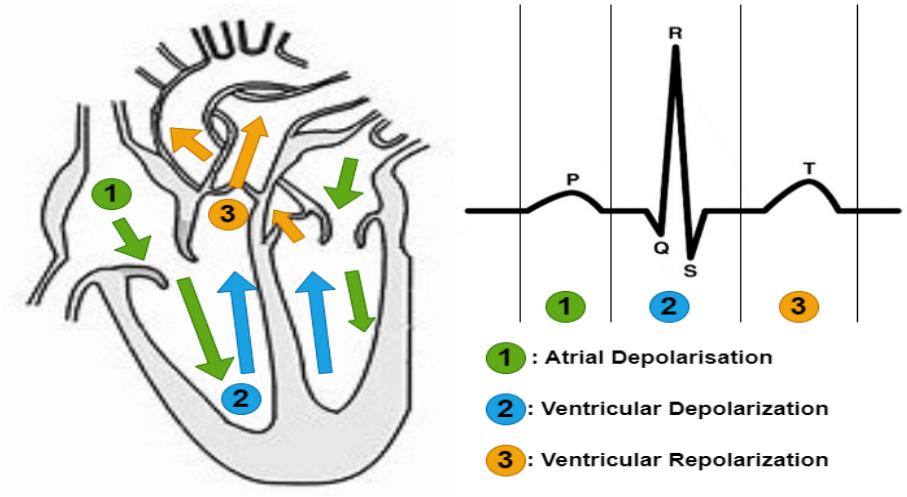

FIGURE 1.3: A single cardiac cycle and ECG signal with the representation of P, Q, R, S and T peaks

1.2 Research Motivation

In the literature review of ECG biometrics, it is possible to identify some challenges that are open to research such as physical activity and emotional status effects on the biometric system, the lack of studies for real-life biometric verification scenarios and the performance of different ECG technologies. Therefore, medical devices, medically approved devices and consumer-based wearable devices with different activity and emotional status conditions were explored in this thesis. To address these issues, the main research questions explored in this thesis are listed below.

1. *What is the baseline equal error rate from various devices using ECG?*

   In order to present a novel biometric verification framework, first of all, the parameters of that biometric model and the baseline of biometric verification performances should be understood. Therefore, the features extracted from ECG waveforms encountered in the literature, new features, different classifiers, various devices, the effects of enrollment and authentication times and different Machine Learning (ML) models in order to examine their real-life applicability indicate a great challenge. To address this challenge, this study offers in-depth research for all of the aforementioned parameters.

2. *Can we accurately detect physical activity or emotional status from ECG waveforms?*

   It is known that different physical activities, various emotional states and different factors (such as medication and alcohol) during the day affect our heart rhythm. Considering the application of ECG-based biometric verification to real life, these changes cause errors in verification performance because the stored template and test data do not match with each other. Investigating this problem and presenting a different way or method other than current research is a subject open to study.

   It is a subject open to research because of the endless possibilities of parameters and models that can be used. In order to examine this problem, different device types and datasets are used in this study. In addition, the performance of the newly introduced features has been tested with the help of different classifiers. Deep Learning (DL) models, a branch of ML, have also been used to investigate the effects of various parameters.

3. *Do physical activity or emotional status classification prior to biometric verification improve performance in medical and wearable ECG devices?*

   This question leads to the proposal of a new biometric verification framework. Although the performance of different devices using today's technology, especially wearable devices, is being investigated, it is an important step for the use of ECG-based biometric verification in access control in daily life.

It is known that ECG signals change with the conditions of daily life and therefore are not used frequently for verification models. In order to find a solution to this problem, ML and DL models and their many different parameters have been investigated.

4. *Do different machine learning models and their parameters contribute to the improvement of device performances?*

   All ML and DL models show different outputs according to the parameters used, datasets, training and testing data sizes etc. For this reason, an in-depth analysis is necessary for the consistency of the model to be created. In this study, parameters such as different classifiers, features, various training and testing data sizes, and different enrollment times in the classical ML model are compared using data from various mobile devices.

   In the DL model, different Convolutional Neural Networks (CNN) structures, optimisers, various epoch numbers and recording times are compared with different image inputs. This study holds significance in observing device performances. It observes the reliability of wearable devices by assessing the influence of diverse ML models and their parameters on the data acquired from medical and wearable devices. In addition, due to the proposed biometric verification framework, we explore which parameters the examined activities and emotional statuses achieve optimum results. Thus, the specific characteristics of each activity or emotional status are examined.

1.3 Thesis Outline

This thesis consists of 7 chapters describing the novel ECG-based biometric verification framework, which utilised the activity classification.

Chapter 2 introduces state-of-the-art studies on ECG biometrics, models that classify activity and emotional status, and a detailed overview of DL models in ECG-based applications. The performance assessments of the biometric models

are given to understand the outcomes of state-of-the-art studies. After stating the related works about ECG biometrics, the open challenges in the literature are indicated for further investigation. In addition, it is stated the proposed biometric verification framework's approach to these challenges.

Chapters 3, 4, 5 and 6 explain the experimental parts of the thesis. Chapter 3 describes the general methodology used in the experiments. The datasets, devices and data collection processes used in this thesis are examined. Then, the data analysis methods we used, pre-processing, feature extraction, exploring a time-frequency representation and explanation of the used classifiers are given.

In Chapter 4, the first phase of our research, the effects of different enrollment periods, attributes and various time-frequency representations on biometric verification models were examined. The performances of various devices have been researched for classical ML and DL models. The effects of the compared parameters on device performances are discussed.

In Chapter 5, the performance of novel features in physical activity classification and emotional status classification are explored and presented. In addition, GoogleNet, ResNet50 and DenseNet201 CNN structures are explained and the effects of DL parameters on activity classification are discussed.

As a final phase in this thesis, the framework of activity-aware biometric verification proposed in Chapter 6 is examined. In the ML model, the performance of samples classified according to their physical activity and emotional state when used in biometric verification models is explained. In addition, the performances of 5 different devices are examined. In the DL section, ResNet50 and DenseNet201 CNN models are tested on wearable and medically-approved devices for different parameters and comparative results are discussed.

In Chapter 7, the last chapter of the thesis, a general summary of our work, answers to our research questions, a roadmap for future work, and conclusions are explained.

1.4 Limitations

COVID-19, defined as a global pandemic by the World Health Organization (WHO) as of March 2020, brought with it a number of restrictions to prevent the spread of the disease. Within the scope of these restrictions, general life including workplaces and universities was disrupted by various restrictions for more than two years. During the 2020 and 2021 years, our access to laboratories, offices and all university facilities is restricted.

Since the beginning of this study, the data collection process and the creation of a new ECG recorder could not be achieved in our study due to the inaccessibility of laboratories, the prohibition of data collection and the inability to have close contact with participants within the COVID-19 restrictions. Due to the limitations in our capacity to collect ECG data, we resorted to utilising datasets that were accessible to the public.

Public datasets require more detailed data pre-processing because the data is not collected directly for your intended use. It is recognised that variables like temperature, the location of the device, and the friction between the device and skin, among others, can influence individual data and generate noise. Moreover, various devices and different experimental conditions contribute to distinct noise levels in the datasets. The pre-processing techniques employed in this study effectively minimized noise and prepared the data for analysis.

Since one of the aims of this study was to compare the performance of medical and wearable devices via the proposed biometric verification framework, the datasets used contributed to our study in terms of device diversity and providing different experimental conditions.

Despite all the setbacks and limited time, we thoroughly examined the proposed biometric verification framework in classical ML models. In addition, we tested the framework using a wearable device and a medically approved device in DL models and proved that the framework was successful in the DL model as well.

Chapter 2

State of the Art

2.1 Introduction

Considering the areas where biometrics is widely used, face, iris and fingerprint recognition models come to mind first. Although the ECG biometric model is very popular in the literature, it is not widely used in the industry. The main reasons are that ECG recording sensors are not as common as cameras and fingerprint sensors, the ECG signal is highly affected by environmental and emotional conditions so the ECG model requires more enrollment time than other biometrics and the lack of long-term stability.

This chapter reviews studies on ECG biometric applications using Machine Learning (ML) and Deep Learning (DL) models. Initially, we will describe the meaning of ML and learning categories. Then, the history and technological development of wearable ECG recording devices will be examined in a separate section. The novel approaches in ECG biometrics will be explained. In this section, classical ML-based ECG biometric identification and verification models will be presented, followed by DL-based models. Due to the extensive study of activity recognition systems in the literature, general models will be introduced first. Subsequently, activity recognition models that use only ECG data will be reviewed. In addition, activity-aware biometric systems will be mentioned. The emotional state

classification will be examined separately from the physical activity classification. Studies examining the effects of different distance metrics will also be reviewed. Finally, by comparing these state-of-the-art studies, the points that need to be researched will be revealed and the contribution of this study to the issues will be stated.

2.2 What is Machine Learning?

Artificial intelligence (AI) is a field of computer science that focuses on creating intelligent machines that can learn and solve problems in a way that is similar to how humans do. This involves developing algorithms and software that can analyse data, recognise patterns, make predictions, and perform other tasks that typically require human intelligence. Some of the key areas of AI research include ML, natural language processing, computer vision, and robotics. With advances in technology, AI is becoming increasingly important in many industries such as healthcare, web search engines and cybersecurity [27]. ML is considered a subset of AI. While AI is the general name for computers or devices that can imitate human intelligence, ML refers to algorithms that perform functions such as making decisions and making predictions. ML is the process through which a computer system acquires intelligence [27]. After the development of neural network models used in ML with backpropagation [28] and CNN [29] algorithms, DL has been considered a subset of ML [30].

DL is an ML algorithm that has a more profound architecture thanks to more layers. Learning performances change as the structures, numbers and orders of the layers in DL change. In this context, different DL models such as AlexNet (for voice biometric authentication [31]), ImageNet (for image classification [32]), GoogleNet (for ECG biometric verification [33]), ResNet [33], and DenseNet (for finger-vein recognition [34]) have been created for the tasks to be obtained from the DL algorithm. In addition, DL architectures are named according to the properties of the layers they contain. For example, while the Deep Neural Networks (DNN)

model has many more hidden layers than Neural Networks (NN) , the CNN model has convolution layers and filters of various sizes, making it more useful in image recognition tasks. The main reason for this is that the CNN model adapts better to 2D and 3D images and distinguishes shape differences in 2D images better than the DNN model [30].

ML can be categorised into four different groups based on their learning approaches: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning [35]. Supervised learning involves training machine learning models with labelled data. These labels have been inspected or assigned by a supervisor. When predicting, labels should describe input features and target outcomes. Supervised learning is generally used for classification problems which involve using an algorithm to classify test data accurately into specific classes and regression problems which are used algorithms for distinguishing dependent and independent variables. Decision Trees (DT), Naive Bayes (NB), k-Nearest Neighbours (KNN) and Linear Discriminant Analysis (LDA) can be given as examples of classifiers used in this type of learning. Fig.2.1 displays the flow chart for supervised learning.

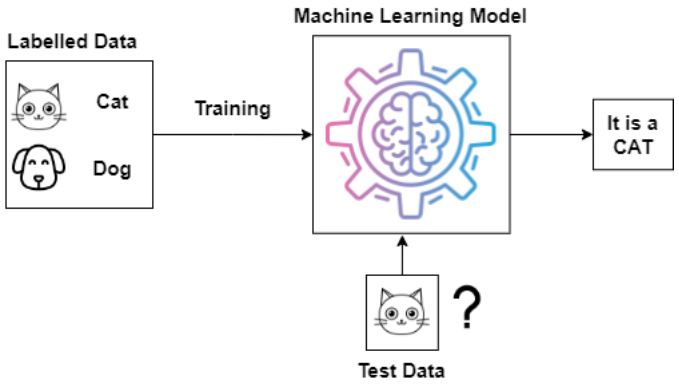

FIGURE 2.1: Supervised learning flow chart

Unsupervised learning is used for the organization of data that has not been previously labelled or classified [35]. This is achieved by identifying similarities within the data set, thereby creating distinct groups. There is no supervisor in this approach. For this reason, the algorithm needs to distinguish different patterns

autonomously. This learning method is used to cluster data, find associations between data, or reduce the dimension of the dataset [36]. Fig.2.2 shows the flow chart for unsupervised learning.

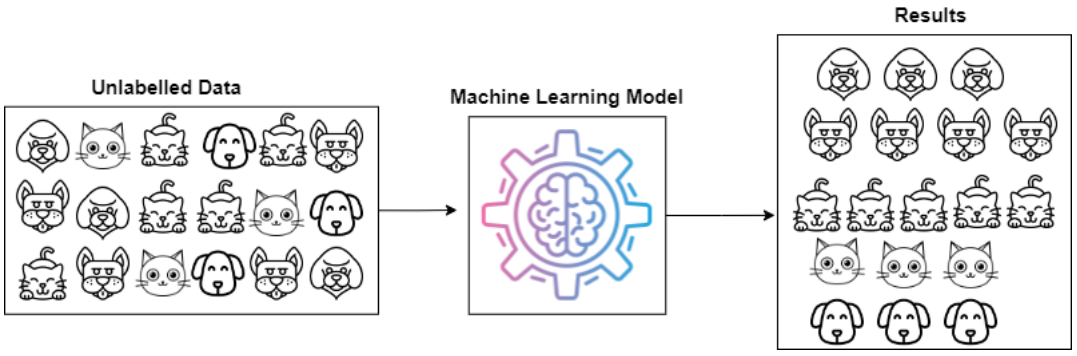

FIGURE 2.2: Unsupervised learning flowchart

K-means clustering (k value is the number of classes or centroids) can be given as examples of clustering models for unsupervised learning. It involves dividing the data into k groups based on how closely the data within each class is related to one another or how close the centres of each class are to each other [37]. Fuzzy clustering, the expectation maximization and self-organizing maps can be mentioned as examples of unsupervised machine learning. Fuzzy clustering has a membership function quantifies the degree to which a specific element is associated with an event. The membership value lies within the interval [0, 1], signifying the extent of influence that the element has in relation to the event [38–41]. The expectation maximization algorithm facilitates parameter estimation in probabilistic models when dealing with incomplete data [42–46]. The self-organizing maps transforms high-dimensional data into a low-dimensional map while maintaining the topological relationships inherent in the original data. Additionally, it automatically organizes the data based on its underlying structures [47–49].

Semi-supervised learning, as the name suggests, is a learning approach that uses both labelled and unlabelled data. Labelled data is in small amounts and unlabelled data is in larger amounts. In this approach, the machine learning model is trained with labelled data. Then, unlabelled data is added to the model and pseudo labels are assigned to these data [50]. When the test data is given to the model, the

test data is classified using these pseudo-labelled and labelled data. Self-training, label propagation and generative adversarial networks are some examples of classifiers in this learning method. One of the common uses of semi-supervised learning is shown in Fig.2.3.

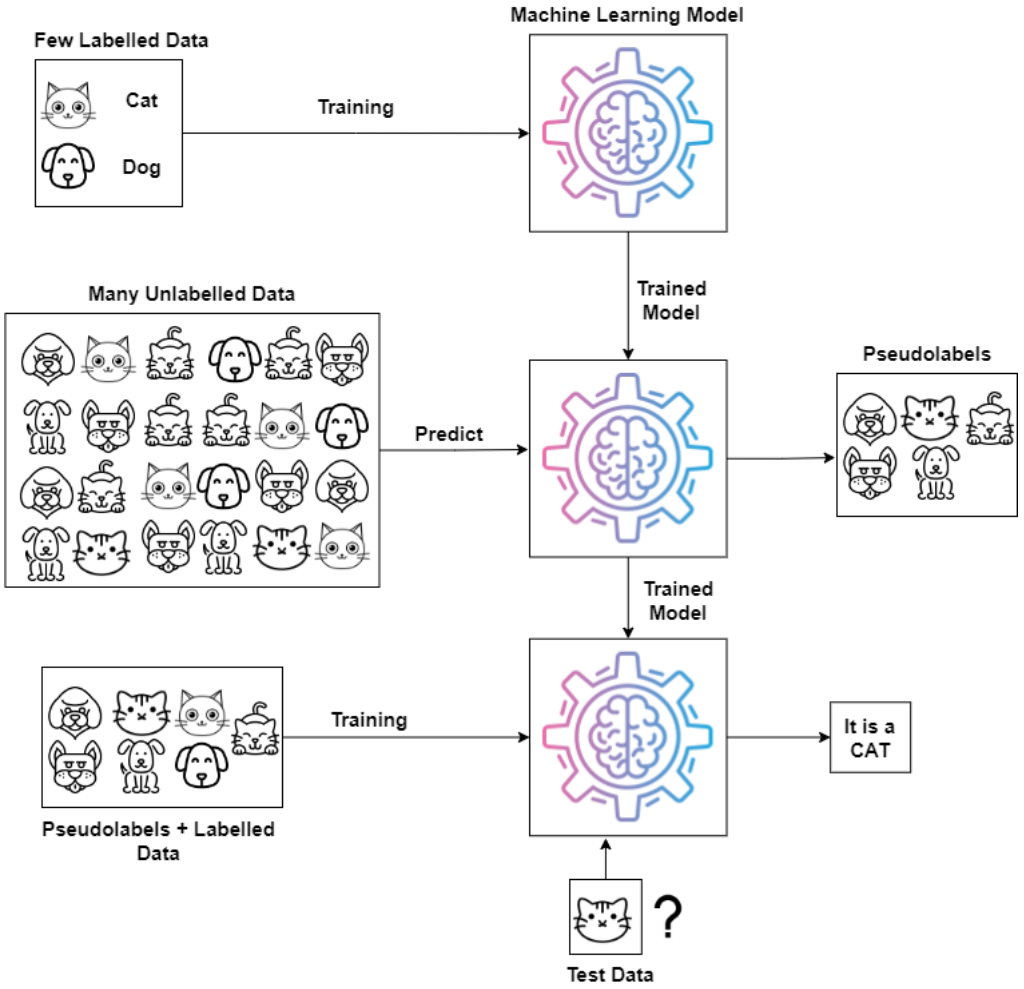

FIGURE 2.3: Semi-supervised learning flowchart

Reinforcement learning, based on the Markovian decision-making process [51], is learning through trial and error, similar to the process of human learning. For instance, a child learns that fire can harm them through direct experience of touching it. Reinforcement learning is also like this. The model makes a mistake and gets feedback about the mistake by observing the environment. This cycle continues until the agent performs the correct behaviour or action and receives the reward. Reinforcement learning agents can be used in examples such as natural language processing, traffic signal controls and image segmentation for healthcare

applications (e.g. diagnosing cancers) [52]. Reinforcement learning is illustrated in Fig.2.4.

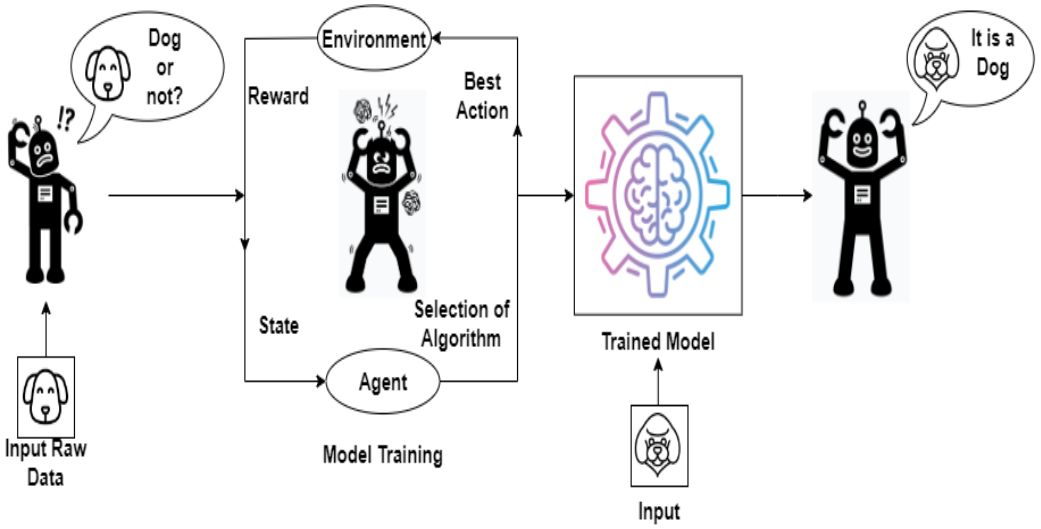

FIGURE 2.4: Reinforcement learning flowchart

In this study, biometric verification, physical activity classification and emotional status classification were investigated using different devices. Due to the presence of labels in the datasets and to see the effect of activities on biometric verification models, various supervised learning models were investigated. This study investigated the performance of wearable devices in biometric verification models under different conditions. Therefore, we will provide a review of the technological advancements in these devices.

2.3 Technological Advancements in Wearable ECG Recorders

The history of wearable devices is quite old but their first use for access control purposes took place in 1998, when financial transactions could be made with the mBarecelet wrist-wearable computer [53]. This technology then continued with the smart jacket in 2000, the wireless headset in 2001, the first smartwatch in 2003, activity trackers in 2007, Google Glass in 2012, and Android clothing in

2014 [53]. Furthermore, combining wearable devices with biometric features for security purposes and using them for access control is a very popular area today [54].

It has been observed that the wearable device market has increased significantly with the use of such wearable devices with health monitoring applications. Using consumer-based wearable devices has been shown to lead people to do more activity [55, 56]. In addition, athletes frequently use wearable devices to actively check their performance, heart rate and health status [57]. These devices are also used in biometric applications to check situations such as the presence of the driver or whether the driver is awake or not [58].

When we examine the development of ECG recording devices in the field of health, we see that 12-lead ECG recording devices are frequently used in the health sector for cardiac health imaging and diagnosing cardiac diseases. However, due to the large number of leads, the establishment of an ECG recording mechanism, and the impossibility of using them daily, these devices can only be used with trained personnel help (e.g. medical doctors and clinicians). To eliminate these difficulties, wearable devices that can be monitored remotely have been produced [59]. Thanks to wearable devices, the number of leads of the ECG recorder has been reduced (generally 1-lead or 2-lead configurations) and it has become consumer-oriented. This has raised the question of whether the performance of consumer-oriented devices is as accomplished as medical devices. Researching this problem, Phan et al.[60] used a pulse oximeter, an LG smartwatch and a 3-lead ECG Powerlab device to measure heart rhythm during sleep. The correlation of the heart rhythm measurement of the smartwatch with other devices was found to be 90% for 4 participants. In this thesis, the answer to this question is investigated for activity classification and biometric verification models.

There are two configurations for recording ECG signals: on-the-person and off-the-person. Electrodes for ECG measurements are placed directly on the human body in an on-the-person configuration system. Conductive materials like gel are generally needed between the body and electrodes for this system to work effectively.

This kind of system is commonly used in medical tools such as bedside 1-lead ECG monitors, attachable patches, and other diagnostic tools [61]. The off-the-person configuration employs devices, having dry-type electrodes, that capture ECG signals without necessitating any specific preparation of the individual with objects or surfaces [61]. This approach is advantageous as it minimizes any potential discomfort or inconvenience to the subject. Examples of these configurations include fingertips and other wearable devices such as smartwatches.

When on-the-person systems are considered to make a diagnosis, it is essential to understand the P, Q, R, S and T peaks and the relationships between them (such as the time difference between two peaks). In this regard, it has been stated that devices such as Zio-XT and CAM patches achieve as successful results as the Holter device [61–63]. Since medical devices are used for diagnostic purposes, they generally limit activities during the day (such as heavy sports training, bathing and even sleeping position) but wearable devices do not have such restrictions. Therefore, it is observed that wearable devices obtain noisier signals and their performance can decrease due to factors such as daily activities, sensitivity to liquids (sweating, showering etc.) and noise caused by friction. Therefore, it is important to investigate the issue of sensitivity to activities. Researching this problem, Muhlsteff et al.[64] created a wearable belt band that can record ECGs and measure activity in patients. ECG recordings were collected from 12 participants for 24-48 hours. It was stated that while 70% of the ECG signals collected during the day were obtained in good quality, this rate was 60% during sleep.

2.4 Biometric Performance Assessments

The efficacy of a biometric authentication model is determined by its ability to accurately predict the correct class in a verification or identification task, using previously unseen data. Verification tasks, which require a subject to confirm their identity, can be conceptualized as a binary classification problem: the system either accepts the data as belonging to the genuine subject or rejects it. Identification

tasks, on the other hand, are multi-class problems in which the biometric system compares the given sample with a list of existing templates and returns the class with the highest likelihood. In both scenarios, if the class prediction is accurate, it is considered a True Positive (TP) if the data was provided by the genuine subject, or a True Negative (TN) in cases of imposter. Conversely, misclassifications are referred to as False Positive (FP) and False Negative (FN).

The ISO is responsible for developing and publishing International Standards. Within this organization, there exists a dedicated Biometrics subcommittee, known as ISO/IEC JTC 1/SC 37, which focuses on the development of standards within the field of biometrics [65]. In March 2022, ISO published a biometric vocabulary, designated as ISO/IEC 2382-37:2022(E) [6]. According to the most recent ISO standards, the terms Equal Error Rate (EER), False Accept Rates (FAR), and False Reject Rates (FRR) are defined as follows:

- FAR represents the proportion of biometric transactions in which false biometric claims are erroneously accepted (i.e. False Accept / Total Imposter Matching).
- FRR represents the proportion of verification transactions in which true biometric claims are erroneously rejected (i.e. False Reject/Total Genuine Matching).
- EER is defined as the value at which the FAR and the FRR are equal.

Metrics such as *Accuracy, FAR* and *FRR* can be derived from the confusion matrix, as illustrated in Table 2.1. These metrics are predicated on direct prediction. Equations (2.1), (2.2) and (2.3) respectively represent the mathematical expressions for calculating the *Accuracy, FAR* and *FRR*.

TABLE 2.1: An example of a confusion matrix

Confusion Matrix		**Responses**	
Total= G+I		**Accept (G)**	**Reject (I)**
Real Condition	**Genuine (G)**	TP	FN
	Imposter (I)	FP	TN

$$Accuracy\,(\%) = 100 \times \frac{TP + TN}{Total} \tag{2.1}$$

$$FAR\,(\%) = 100 \times \frac{FP}{(TP + FP + TN + FN)} \tag{2.2}$$

$$FRR\,(\%) = 100 \times \frac{FN}{(TP + FP + TN + FN)} \tag{2.3}$$

Biometric systems generate a probability that a given sample belongs to a particular class, typically expressed as a similarity or dissimilarity score between the sample and a template. The prediction of the class is based on this score and a threshold that minimizes classification error.

In an ideal situation, a classifier would be capable of perfectly distinguishing between two distributions. However, in practice, there is often overlap between the distributions. The selection of a threshold value will determine the classification errors that arise from the tails of the distributions crossing it. Type 1 errors (FP) occur in the imposter distribution, while Type 2 errors (FN) occur in the genuine distribution. The optimal threshold is one that minimizes errors and is typically located at the point where the FAR and FRR are equal. This point which is referred to as the EER is commonly found at the intersection of the two distributions [66].

2.5 Novel Approaches in ECG Biometrics

In order to identify current and novel research on this thesis, our literature review generally focuses on the year 2016 and later. In this section, we undertake a comprehensive review of the relevant academic literature, which spans several crucial domains of study. These include the utilisation of ECG data for biometric verification and identification, the classification of physical activities, the exploration of activity-aware biometric systems, as well as the classification of emotion status and distance metrics comparisons. Each of these areas contributes significantly

to the context of our research and provides a foundation upon which our study is built.

2.5.1 ECG biometric verification and identification

ECG signals have been used for biometric identification and verification purposes for two decades via ML and DL models. ML is a process that utilises features extracted from data, which represent the raw data [15]. On the other hand, DL models are a more advanced version of traditional networks, which contain several hidden layers. Features are calculated by deep networks. In that way, the learning model finds optimal features and uses these features for the learning process. Additionally, DL models can automatically identify previously unseen image features from original images, without any human intervention or manual feature extraction.

In ML studies, there are two feature categories: fiducial and non-fiducial features. Fiducial features include P, Q, R, S and T peaks, time differences between two peaks and the characteristics of these peaks, etc. Many studies in the literature, including ours, have investigated the effects of features on ML [67–69]. Non-fiducial features are obtained by applying time analysis or frequency analysis to the ECG signal. Some of these techniques are Auto-Correlation (AC) [70], wavelet transformation [71], Dynamic Time Wrapping (DTW) [61] and Discrete Cosine Transformation (DCT) [72].

Ingale et al.[61] conducted a comparative analysis of ECG-based biometric authentication using six different ECG datasets from various devices. They employed extended Kalman filters and Infinite Impulse Response (IIR) filters in conjunction with ECG-beat and RR interval segmentations. They extracted 30 fiducial and non-fiducial features from the ECG signals, resulting in an Equal Error Rate (EER) ranging from 0.5% to 7% for fiducial features, and a broader range of EERs for non-fiducial features. For the biometric verification process, they used Euclidean distance and DTW methods as matching algorithms. They found that

datasets collected by the Biopac MP36 2-channel ECG recorder, a 16-lead medical device, and 2-lead palm and finger ECG sensors had relatively lower EERs compared to the ECG-ID dataset (comprising single-lead healthy ECG recordings) and the MIT-BIH dataset (containing medical ECG recordings from both healthy and unhealthy subjects). They concluded that fixed window data segmentation, fiducial features, and IIR filter yielded better EERs than non-fiducial features, Kalman filter, and RR interval segmentation methods. The study included both healthy and unhealthy subjects but did not consider different activity cases or devices recording simultaneously [2].

Choi et al.[69] collected ECG signals for 60 seconds from 175 subjects using a 2-lead mobile sensor. They extracted eight fiducial features and used various classifiers, including Support Vector Machines (SVM), Simple Logistic, NB, Random Forest (RFC), AdaBoost, Multi-Layer Perceptron (MLP), Bayes Net and RBF Kernel-based SVM for both single-beat and multi-beat authentication scenarios. In the single-beat verification scenario, the EERs ranged from 4.46% to 9.51% depending on the classifier used, with SVM performing the best and NB performing the worst. EER values were stated as 4.46% SVM, 8.81% simple logistic, 9.51% NB, 6.23% RF, 6.51% Adaboost, 6.72% Bagging, 8.54% MLP, 6.10% Bayes Net and 9.3% RBF network, respectively. However, they expressed that more reliable results were obtained when more beats were used for biometric authentication. Therefore, they tested their models using the best-performing SVM classifier with data ranging from 3 to 15 seconds. The system's EERs decreased from 8% (for 3 seconds of data) to 1.87% (for 15 seconds of data), indicating that increasing the testing time reduces the EER [2]. Although their results are quite successful, they differ from our study because they do not use any activity or multi-session data. In addition, they explained that the single-beat biometric authentication model they implemented did not provide reliable results and that they obtained quite different EERs from subject to subject. In the case of actual verification scenario, the number of genuine or imposter subjects was not specified. It was stated that the data of 127 subjects were used for validation, but no number of subjects was specified for authentication.

In a study by Krasteva et al.[68], they analysed 12-lead ECG signals from 460 patients in a model designed for verification. They utilised 8 fiducial features from each of the 12 leads and discovered that the most effective results were derived from the frontal plane leads (I, -aVR or II). Each individual provided two sets of ECG recordings, each lasting 10 seconds, while at rest. The data was split evenly for training, with an equal number of genuine and imposter samples, while the remaining data was used for testing, with a higher number of imposter samples. The LDA classifier was used for biometric verification. Depending on the configuration of the ECG leads, they reported EERs ranging from 3.7% to 32.4%. In a study conducted by Pavia et al.[67], they utilised a dataset of single-beat ECGs and extracted fiducial points for the task of identification. They experimented with different training durations using the SVM classifier and discovered that the highest identification accuracy achieved was 97.5%.

ECG recordings taken on different days can show variations due to changes in activity levels and emotional states [73, 74]. Lehmann et al. [74] collected ECG recordings from 20 subjects over a week using a chest band, with a sampling frequency of 1024 Hz. For their study, they downsampled the data from 1024 Hz to 256 Hz according to the literature recommendations [75, 76]. They used the peak-to-peak intervals of QR, RS, and QS as features and applied RF, NN, and SVM classifiers. Their study utilised 2000 samples for training and 500 samples for testing, for both genuine and imposter cases. When considering data collected over different days, they have 3 scenarios. In the first scenario, they used a single-day data for training and testing (i.e. single-day data for enrollment). They found 21.91% EER with RFC, 26.17% with NN and 28.08% with the SVM classifier. In the second scenario, they used 2 days data for enrollment and they achieved 20.80% with RFC, 21.02% with NN and 28.10% with SVM. In the last scenario, they used 3 days data for enrollment and they obtained 19.54% with RFC, 20.16% with NN and 26.77% with SVM. The best results were achieved when data from 3 days were used for training, while the highest EERs were observed when only 1 day of data was used for training.

Patro et al. [77] compared various feature selection methods for biometric identification using ECG-ID and PTB datasets. They used ECG recordings of 20 seconds per person and extracted 72 QRS-based features. To reduce the number of features, they employed wrapper feature selection and embedded feature selection methods. In most instances, RF had better accuracy rates than SVM. When using all 72 features for direct classification, they achieved accuracy rates between 89% and 92%. The wrapper feature selection method yielded the best results, with accuracy rates ranging from 91.54% to 95.30%. Meanwhile, the embedded feature selection method resulted in accuracy rates between 92.2% and 94.90%. Thus, it was proven that using more features does not always result in higher accuracy rates.

When we examine DL studies, Kim et al. [73] proposed a method for biometric authentication using short ECG recordings in the ECG-ID dataset. While 83 subjects were in the first season, only 25 subjects out of 83 were in the second-day season. They employed QT interval correction with the Sum of Squared Difference (SSD), Gated Recurrent Unit (GRU), and AlexNet 1D CNN DL methods. They created 1-pulse window and 3-pulse window sizes from ECG signals to use in authentication. Each subject has 12-pulse ECG data per recording. Even if they stated that it is the authentication study, the number of imposters, genuine samples and unseen subjects in the authentication step is not specified. In addition, they only mentioned accuracy, precision, sensitivity and specificity but we calculated EERs using $\frac{(1-Sensitivity)+(1-Specificity)}{2}$ formula for easy comparison. In one scenario, they used data from two different days for both training (1 record each from 83 subjects and 25 subjects) and testing (1 record each from 83 subjects and 25 subjects, with 4-fold cross-validation)[2]. When they used 1-pulse for authentication, they achieved EERs of 13.58%, 1.85%, and 2.25% for the SSD, GRU, and CNN methods, respectively. Although a 1-pulse scenario is ideal for faster authentication, it is poor overall system reliability because it may over-fit and not reflect all the data. When 3 pulses were used in authentication, the EERs were 9.78%, 2.82%, and 2.29% respectively for the same methods. In another scenario, they

used 1 record from 83 subjects for training and 2 records from 25 subjects with 2-fold cross-validation for testing data from different days. The EERs obtained were 12.44%, 2.15%, and 3.81% for SSD, GRU, and CNN methods respectively when using 1-pulse, and 9.43%, 3.66%, and 3.1% EERs when using 3 pulses. Despite achieving significant results, there was a substantial difference between sensitivity and specificity, indicating that the number of genuine and imposter samples may not be equal [2].

Using five different datasets, Li et al. [78] compared biometric identification performances with the cascaded CNN model. 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). Healthy and unhealthy subjects were used together in their study. All data were equalized at a 250 Hz sampling frequency. R-peaks segments were used to feed the CNNs. They used F-CNN, M-CNN and Cascade CNN and found 97.8%, 98.1% and 99.3% identification accuracy rates respectively. They expressed that F-CNN is useful for multi-class classification. However, F-CNN can be easily affected by data variance. For this reason, they used the same dataset, not multiple or merged datasets. M-CNN was used with raw ECG signals. When they fed the M-CNN with R-peaks segments, the results were lower than raw signals. F-CNN was used to learn features and these features were used to feed M-CNN. In this way, they achieved higher identification results.

AlDuwaile and Islam [79] also used R-peaks which were calculated from a 0.5-sec 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. The longest training time (132 min) came from GoogleNet because of the high number of (5.9 million) learnable parameters and a large number of layers (144 layers). The ResNet training time (48 min) was the smallest one among other CNN structures. The number of learnable parameters (4.8 million) and layers (71

layers) was smaller than GoogleNet. 100 subjects from PTB and 90 subjects from ECG-ID datasets were used in their study. In blind segmentation, 0.5 sec, 1 sec, 1.5 sec, 2 sec, 2.5 sec and 3 sec time windows were used and the 2 sec time window was selected as the best-performed window size for the biometric identification. In this case, GoogleNet achieved the best accuracy rate of 98.14%. In heartbeat segmentation, 0.5-sec, 0.75-sec and 1-sec were used with 100 records. 0.5 sec time window (single heartbeat) heartbeat segmented data have higher accuracy rates than other cases for biometric identification. ResNet, MobileNet and GoogleNet achieved 100%, 100% and 99.90% accuracy rates, respectively. In addition, they obtained biometric verification results between 2.5% and 5.8% the half total of EER (i.e. (FRR+FAR)/2). The lowest half total of EER was achieved by ResNet and GoogleNet however, they did not mention any imposter samples and real EERs.

Begum et al. [80] compared 4 distinct DenseNet CNN architectures with several train/test sizes on 8 ECG datasets (MITDB, NSRDM, PTBDB, QTDB, ECG-ID, IAFDB, CUDB and MIMIC-II/III) for biometric identification purposes. These datasets include healthy and unhealthy subjects with a range of subject sizes. They equalized the sampling frequency at 500 Hz for all datasets and each 1000 samples were selected to create template images. The system which was used for the data of 250 subjects had higher identification accuracy than 120 subjects. The test prediction accuracies for four different architectures were as follows: Architecture #1, which had 5 convolutional layers, 3 concatenation layers, and 3 filters in each convolutional layer, achieved an accuracy of 99.42%. Architecture #2, with the same number of convolutional and concatenation layers but with 16 and 5 filters in each convolutional layer, achieved a higher accuracy of 99.84%. Architecture #3, which had 6 convolutional layers, 4 concatenation layers, and 10 filters in each convolutional layer, had an accuracy of 99.8%. Lastly, Architecture #4, with 6 convolutional layers, 5 concatenation layers, and 10 filters in each convolutional layer, achieved the highest accuracy of 99.94%. It was also observed that the best identification rate was achieved when using 90% of the data for training.

Hammad et al. [81] used 5 distinct ResNet-attention CNN architectures for biometric authentication. They used 2-second time windows on two datasets: PTB (290 subjects) and CYBHI (63 subjects). The raw ECG data was processed using ResNet. During training, they used a batch size of 10 samples across 100 epochs. The data was divided into 70% for training, 20% for validation, and 10% for testing. They employed a 10-fold cross-validation method for biometric authentication. The number of genuine and imposter samples was not the same: in the PTB dataset, they used 300 genuine and 199 imposter samples, while in the CYBHi dataset, they used 240 genuine and 120 imposter samples. The EERs were calculated as 1.53% for PTB and 0.27% for CYBHi datasets using the 1-D CNN approach. However, with the ResNet-attention approach, the EERs were slightly different: 1.39% for PTB and 0.68% for CYBHi datasets.

Rahiem et al. [82] used the MWM-HIT dataset for ECG-based authentication. They 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. They achieved 99.4% authentication accuracy from SVM and 99% from the KNN classifier. In the dataset, 10 seconds 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. Byeon et al. [33] used PTB-ECG (290 healthy and unhealthy subjects, 12-lead configuration, various ECG recording times from 23 seconds to 2 minutes) and CU-ECG (100 subjects, 10-second ECG recordings, single-lead configuration). They used each R-peak's scalogram images to feed AlexNet, GoogleNet and ResNet CNN structures. They used 50% training (0.9 or 0.75 training and 0.1 or 0.25 validation) and 50% testing data. They found 0% EER, 100% accuracy rate, 0% FRR and FAR with GoogleNet and 0% EER, 99.17% accuracy rate, 2% FAR and 0% FRR with ResNet. 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 addition, Byeon et al. [83] 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. They found the best identification rates when they used Xception CNN and MFCC images.

Pereira et al. [84] used the CYBHi dataset for identification and verification purposes. They used 63 subject data with scalogram images and each cardiac cycle image in a 15-layer CNN. For biometric verification, Manhattan distances between templates were used as a matching algorithm. In addition, while the original scalogram image size was 224x224, they reduced the size to 56x56 using the ICA method and examined the performance difference between these two images. They could not obtain a clear conclusion as to whether dimension reduction is advantageous or not. They also observed that cardiac cycle images gave better verification results than scalogram images. In the study where EER results were not shared, it was stated that 57 of 63 people were confirmed as genuine users and 54 of them were verified as imposter users in the optimal performance of the cardiac cycle. It was also stated that in the best performance of the scalogram image, 62 out of 63 people were verified to be genuine users and 53 were imposter users.

Ciocoiu et al. [85] 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 CNN (2-D CNN, AlexNet, SqueezeNet, GoogleNet, VGG16, MobileNetv2, Inceptionv3, ResNet50, DenseNet201 and Xception) models. ResNet50 achieved the highest identification rates. In addition, they used their own 2-D CNN model for verification. They used identification to create templates and Euclidean distance for matching. 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. They obtained 9.69%

ERR when using a single template, 5.48% EER for 3 templates, 4.86% EER for 5 templates, and a limit of 4.4% EER for more than 13 templates. 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 Table 2.2, there are further studies that used different datasets and models.

TABLE 2.2: Table of recent biometric authentication studies

Authors	Datasets	# of Subjects	Recordings	Methods	Performances
Kang et al. [86] (2016)	Private (wearable/watch)	28	30 s registration, 3 s authentication	Cross-correlation of ECG beats	FAR: 5.2% FRR: 1.9%
Wieclaw et al. [87] (2017)	Private Lviv (fingertips)	18	137 records x 10 s, 88 records in training, 49 records in testing	MLP	Identification accuracy: 88.97%
Luz et al. [88] (2018)	CYBHi UofTDB	65 100	CYBHi: 21886 Heartbeats UofTDB: 61312 Heartbeats Session 1: 50%, Session 2: 50%	Raw ECG- 1D CN + Spectogram -2D CN	CYBHi same session EER: 1.33% CYBHi 2 sessions EER: 12.78%, 13.93% UofTDB 2 sessions EER: 14.27%
Yin et al. [16] (2019)	In-house	645	~30 s registration 4, 3 and 2 ECG beats authentication	NN feature extraction and Cosine similarity matching	4 ECG beats EER: 1.1% 3 ECG beats EER: 2.0% 2 ECG beats EER: 3.9%
Melzi et al. [11] (2023)	In-house PTB ECG-ID CYBHi	645 165 90 65	5162 genuine/25801 imposter 113 genuine/565 imposter 89 genuine/445 imposter 63 genuine/ 315 imposter	Single-lead CNN and Euclidean distance matching	5.55% EER 5.12% EER 0.26% EER 5.44% EER
Prakash et al. [89] (2023)	ECG-ID	90	20 beats per person	Siamese Network	Single beat image: 91% accuracy Dual beat image: 99.85% accuracy Triple beat image: 99.90% accuracy
Ours Phase 1 [1]	E-HOL WeSAD	200 15	50, 150, 250, 500 seconds enrollment (5,15,25,50 genuine samples), 10 seconds verification (1 genuine sample)	13 fiducial features LDA, DT, NB and KNN	**E-HOL:** NB: 3.64% - 6.30% EER DT: 7.89% - 27.91% EER LDA: 15.94% - 37.33% EER KNN: 32.01% - 45.67% EER **WeSAD:** NB: 3.02% - 4.57% EER DT: 1.60% - 31.88% EER LDA: 9.97% - 32.03% EER KNN: 31.94% - 48.79% EER
Ours Phase 1 [2]	Vollmer WeSAD	13 15	Several genuine and imposter samples (cf. Table 4.3)	4 seconds time window Manhattan and Euclidean-based features DT and NB	**NB:** RespiBAN: 19.81% - 30.67% EER Faros: 17.36% - 25.89% EER SomnoTouch: 19.54% - 27.62% EER Nexus-10: 18.62% - 25.31% EER Hexoskin: 18.27% - 26.04% EER **DT:** RespiBAN: 26.33% - 36.43% EER Faros: 25.16% -32.19% EER SomnoTouch: 25.88% - 33.51% EER Nexus-10: 24.87% - 30.73% EER Hexoskin: 25.42% - 31.67% EER
Ours Phase 1 DL	Vollmer	13	Several genuine and imposter samples (cf. Table 4.4)	2, 4 and 10 seconds time windows 1,3 and 5 genuine samples for enrollment ResNet50 and DenseNet201	**ResNet50:** Faros: 11.15% - 39.64% EER Hexoskin: 11.21% - 41.48% EER **DenseNet201:** Faros: 12.81% - 40.81% EER Hexoskin: 8.88% - 41.06% EER

Some cryptographic methods are used to ensure the security of biometric models. In some studies, attack protection methods have been compared, especially for ECG biometrics that are exposed to replay attacks. Camara et al. [90], who compared three methods of cryptographic encryption using the randomness of ECG signals, used the E-HOL-03-0202-003 (E-HOL) dataset. The entropy of ECG data from healthy subjects is lower than that of data from unhealthy subjects. For this reason, they stated that it is a more difficult case in terms of cryptographic encryption. They stated that in this case, precautions can be taken against attacks by using the True Random Number Generators method. By comparing many key generation methods, González et al. [91] proved that time-invariant keys can be produced using the E-HOL dataset, different keys can be generated between users, and that these keys are difficult to reproduce. Therefore, they proved the uniqueness, time-invariant and invulnerable structure of ECG signals. In addition, they stated that data classified according to different activities would be better in key generation and protection against attacks.

Although the results obtained and the methods used in previous studies are noteworthy, it has been observed that datasets with long recording periods are less preferred for biometric verification applications. In addition, in machine learning applications, the effect of different feature sets, different ML and DL parameters and different types of devices on the results has been observed. When considering real-life conditions, ECG data generally taken from wearable devices at different times of the day are used for biometric verification purposes. However, no studies in the literature investigate different physical activities or emotional states for different devices for biometric verification purposes. Our study contributes to the literature by exploring these issues.

2.5.2 Physical activity classification and activity-aware biometric systems

Some studies have investigated activity recognition using mobile and wearable devices before biometric user identification and verification [70, 75, 92–94]. However, in addition to ECG signals, most also use gyroscope or accelerometer data for activity classification. For instance, Mekruksavanich et al. [95] used gyroscope and accelerometer data to recognise 12 daily activities in the USC HAD dataset and 6 activities in the UCI HAR dataset with CNN and Long Short-Term Memory (LSTM) DL structures. The activities in the dataset were divided into two categories: Dynamic activities such as walking, jumping, and running, and static activities such as sitting, lying, and standing. The dataset consisted of 30 subjects in the UCI HAR dataset and 14 subjects in the USC HAD dataset. The ConvLSTM algorithm achieved an accuracy rate of 91.24% in activity classification using the UCI HAR dataset and the CNN-LSTM algorithm yielded an accuracy rate of 87.77% using the USC HAD dataset. The study found that dynamic activities (walking-related activities) were classified more accurately than static activities (posture-related activities). The highest mean user identification rates were obtained using walking-related activity data, with 91.77% from the UCI HAR dataset and 92.43% from the USC HAD dataset [2].

Batool et al. [96] conducted a study in which they used gyroscope data from an IoT device to recognize three daily activities (walking, sitting, and standing) and biometric identification separately. The main objective of the study was to demonstrate that IoT data can be utilised with the RFC classifier for activity classification and biometric identification. They achieved an accuracy rate of 93% in activity recognition and between 81% to 100% accuracy rates in biometric user identification, using 10-fold cross-validations [2].

Butt et al. [97] used scalogram images with AlexNet and GoogleNet CNNs to classify falling, daily activities and resting. They used ECG signals from the wearable device. They collected data from 8 subjects and compared initial learning rates, 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 (the best performance is 99.2% accuracy rate) had higher results than GoogleNet (the best performance is 98.4%).

Cosoli et al. [98] used the chest-worn (Zephyr BioHarness 3.0) device and smartwatch (Samsung Galaxy Watch3) performances for ECG-based activity classification. They collected resting, walking, slow running and running data for 30 seconds duration each from 30 healthy subjects. They extracted Heart Rate (HR)-based features (such as mean and standard deviation of HR). They compared both ML and DL classifiers with 70% training and 30% testing data. While the RF classifier achieves the highest classification accuracy rate of 81%, it is followed by LSTM at 79%, SVM at 78%, NB at 76%, Simple Logistic at 70% and DT at 66%.

Liu et al. [94] used ECG signals and accelerometer data that were collected from a wearable chest sensor from 13 subjects. Finite Impulse Response (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% 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.

Kim et al. [93] 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. During the enrollment and testing stages, 125 beats were used each time. They found EERs from 0% to 5.61% with SWT features, from 1.77% to 35.38% with the DWT method, from 0.42% to 36.55% with the short-time Fourier transform method and

from 0.94% to 25.74% with the AC/LDA method. In addition, they addressed that sitting and standing activities have lower EERs than the exercise activity.

In a study by Martin et al.[99], they analysed ECG signals from 55 subjects and extracted features from the QRS complexes using Non-Differentiation (ND), First Differentiation (FD), and Second Differentiation (SD). They investigated the impact of different enrollment proportions (25%, 40%, 50%, 60%, 75%, 100%) on EERs during sitting, resting, and exercising, using a MLP classifier. They observed that for sitting and resting activities, the EER decreased as the enrollment proportion increased, but this trend was not seen for exercise. Furthermore, the lowest EERs were obtained from the resting activity, while the highest EERs were observed during exercise. When 100 QRS complexes were used in enrollment (i.e. short enrollment case), they found EERs ranging from 2.79% to 4.95% for ND, FD, and SD. However, when 187 QRS complexes were used in enrollment (i.e., long enrollment case), the EERs ranged from 2.69% to 4.71%. Although it has been reported that different activities and data collection seasons affect performance, this study does not include any activity classification. In Table 2.3, there is a summary of state-of-the-art studies in activity-aware biometric models.

For the biometric authentication task, Nawawi et al. [100] used the Hexoskin wearable device to collect data from 11 subjects during their walking, standing and sitting activities. 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. The data of a single person was used as genuine and the data of the remaining 10 people was used as imposter. 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.

TABLE 2.3: State-of-the-art studies in activity-aware biometric models

Authors	Datasets	# of Subjects	Recordings	Methods	Performances
Wahabi et al. (2014) [70]	UofTDB	1020	69 subjects 50% training / 50% testing Sitting,standing,exercise,supine and tripod activities	DWT features, Time-frequency content, EigenPulse and AC/LDA methods	Accuracy rates in activity classification: 52% - 96% EERs: 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) [92]	UofTDB	1020	52 subjects 50% training / 50% testing Sitting,standing,exercise,supine and tripod activities	DWT features AC/LDA and SVM classifiers	Posture classification accuracy rates Tripod: 94.12% Sitting, standing, resting: 98.04% Mean EER: Same posture: 1.50% Different postures: 8.24%
Batool et al. (2017) [96]	Private	19	Accelerometer and ECG Total: 8 hours of activities 2 hours with each subject. Standing, sitting and walking activities	10 fold cross-validation RFC classifier 50% training / 50% testing, 1 second time window with 150ms overlapping	Activity classification: 93.13% accuracy rate Biometric identification: 85.34% accuracy rate
Liu et al. (2018) [94]	Private	13	Accelerometer and ECG Total: 134.1 min Standing, sitting, squatting, coughing and walking activities	64 sample windows with 32 samples overlapping (each window is 1.28 sec), Frequency and Time domain features	Decision Trees (DTs): Mean 3.08% DER
Kim et al. (2019) [93]	Private	104	70 sec x 5 trials (at least 250 beats per person) Sitting,standing and exercise activities	SWT and Inf-FS features, 125 beats for enrollment and 125 beats for testing, DWT, STFT and AC/LDA methods	EERs: SWT: 0% - 5.61% DWT: 1.77% - 35.38% STFT: 0.42% - 36.55% AC/LDA: 0.94% - 25.74%
Butt et al. (2021) [97]	Private	8	Varied recording time 80% training / 20% testing and 60% training / 40% testing Laying, rolling, falling down and daily activities	Scalogram images, AlexNet and GoogleNet CNN	Best accuracy rates: AlexNet: 99.2% GoogleNet: 98.4%
Mekruksavanich et al. (2021) [95]	UCI HAR (6 activities) USC HAD (12 activities)	30 14	Accelerometer, Gyroscope, Magnetometer Training window length: 128 70% training/ 30% testing	1D CNN, LSTM CNN-LSTM, and ConvLSTM Ensemble classifier for biometric identification	Activity classification accuracy rates: UCI HAR: CNN: 90.322%, LSTM: 90.278%, CNN-LSTM: 90.356%, ConvLSTM: 91.235% USC HAD: CNN: 85.656%, LSTM: 83.112%, CNN-LSTM: 87.773%, ConvLSTM: 85.571% Biometric identification accuracy rates: UCI HAR: 91.78%, USC HAD: 92.43%
Cosoli et al. (2023) [98]	Private	30	30 sec x 4 activities 70% training/ 30% testing Resting, walking, slow running and running	ML with HR-based features and DL classifiers	Activity accuracy rate: RFC: 81%, LSTM: 79% SVM: 78%, NB: 76% Simple Logistic: 70%, DTs: 66%
Nawawi et al. (2023) [100]	Private	11	15 min (80% training/ 20% testing) Standing, sitting and walking activities	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
Ours (Phases 2 and 3) (ML) [2]	Vollmer (resting,walking, standing and uphill walking)	13	6-min per activity 80% Training/20% Testing, 60% Training/50% Testing and 50% Training/20% Testing sample cases	4-s time windows Manhattan and Euclidean distance features KNN,SVM, DT, BT, S-KNN and NN DT and NB for biometric verification	The best activity classification accuracy rates: S-KNN Faros: 97.05% - 91.33% SomnoTouch: 96.88% -91.35% Nexus-10: 97.94% - 90.79% Hexoskin: 96.94%- 92.35% Min/Max EERs: Faros: 6.31%- 13.89% SomnoTouch: 6.67%- 14.24% Nexus-10: 6.36%- 13.89% Hexoskin: 6.31%- 14.38%
Ours (Phases 2 and 3) (DL)	Vollmer	13	Several genuine and imposter samples (cf. Tables 5.7 and 6.6)	2, 4 and 10 seconds time windows Scalogram, Spectrogram and Mel-spectrogram images 10, 20 epochs and ADAM, SGDM optimizers 1,3 and 5 genuine samples for enrollment GoogleNet, ResNet50 and DenseNet201	**ResNet50:** Faros: 11.15% - 39.64% EER, 45.96%- 68.55% accuracy rates Hexoskin: 11.21% - 41.48% EER, 46.97%- 65.18% accuracy rates **DenseNet201:** Faros: 12.81% - 40.81% EER, 47.73%- 69.85% accuracy rates Hexoskin: 8.88% - 41.06% EER, 43.94%- 65.79% accuracy rates **GoogleNet:** Faros: 25%- 70.81% accuracy rates Hexoskin: 25%- 63.55% accuracy rates

Wahabi et al. [92] 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 (50% training, 50% testing). Biometric verification was performed using Discrete Wavelet Transform (DWT) feature extraction, AC/LDA method and SVM classifier. In posture classification, the tripod position achieved a 94.12% accuracy rate, while the other positions achieved a 98.04% accuracy rate. 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. [70] added an *exercise* activity to the existing postures. In the study [70], 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. For this reason, it was observed that there were not equal numbers of subjects or samples for each activity. Half of the data is reserved for training and the other half for testing. EERs obtained in the DWT method range from 2.62% to 32.87%. EERs obtained in the Time-frequency content method range from 1.58% to 33.82%.EERs obtained in the EigenPulse method range from 11.06% to 39.74%. EERs obtained in the AC/LDA method range from 1.44% to 24.10%. Activity classification results also vary from 52% to 96% accuracy rates for different methods. This study stated 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). These studies form the basis of our proposed framework, using fingertip ECG data (mobile sensor) and comparing different methods [70, 92]. 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, and posture classification and biometric verification are examined separately.

2.5.3 Emotion classification and distance metrics comparisons

Classification of emotional statuses is a very common topic in the literature. Since it has been proven that ECG signals are affected by emotional state changes [2], it is quite common to use ECG in emotional status classification. Although some studies have examined both emotional status classification and biometric identification models using ECG signals [101, 102], the emotional status-aware biometric verification model is a subject open to examination in the literature.

It was stated by Gupta et al. [103] that people's fear or anxiety levels should be determined and they should be allowed to withdraw money from ATMs according to their emotional state. A biometric framework has been presented in which a person who is determined to be in fear/anxiety based on ECG signals cannot withdraw money. It has been stated that ATM security will be increased in this way. Although this idea was mentioned in 2018, no study was conducted and the results of the research were not stated. Our proposed biometric verification framework is the first study that shows similarities with this idea.

Bras et al. [101] showed that emotional status classification and biometric identification can be achieved using ECG signals. They used the ECG signals of 25 subjects from the Emotional dataset to classify neutral, disgusted, and fearful emotions (25-min recordings per emotion) and to recognise people in these emotion classes. They trained the KNN (k=1) classifier with 420 seconds of ECG from each emotional state and tested it with all the remaining data. In the biometric identification model, the test results obtained a 98.5% accuracy rate for the neutral emotional state, while a 97.2% accuracy rate was obtained for the other two emotional states. In the emotional status classification model, the accuracy rate of the test data was stated as 77% for fear, 74% for neutral emotion and 63% for disgust.

Li et al. [102] compared many deep learning parameters using 4 different ECG datasets: Dreamer, Decaf, ECG-ID and HCILAB. While some of these data sets

include emotional status, some only contain ECG data. In this study, 9 different emotional states were divided into different training and test datasets and classified with 6 different deep learning models (1D CNN, Bi-LSTM, GRU, DNN, Collaborative-Set Measurement and set-based distance measure). Additionally, 6 distance metrics (Euclidean, Mahalanobis, Learnt Mahalanobis, Manhattan, Chebyshev and Cosine) were compared. However, they could not state any clear conclusion because their results did not show a consistent trend.

In other studies using the Multimodal Dataset for Wearable Stress and Affect Detection (WeSAD) dataset, RFC, LDA [104], DT, AdaBoost, KNN [105], SVM [106], Multimodal-Multisensory Sequential Fusion (MMSF) [107], Self-supervised representation learning [108] and CNN [109, 110] methods were used for emotional state classification.

When we examine classical ML models, Vaz et al. [104] used ECG, Electrodermal activity (EDA) and Electromyography (EMG) signals to extract 109 features but they used 55 features for neutral and anxiety classification tasks. They split the data into 80% training and 20% validation. They obtained a minimum accuracy rate of 62.7% from the LDA classifier and a maximum accuracy rate of 92% from the RFC classifier. When considered for binary classification, a maximum performance of 92% was achieved despite the use of 3 signals together and 55 features. Garg et al. [106] used ECG, body temperatures, respiration, EMG and EDA signals with RFC, SVM, KNN, LDA and AdaBoost classifiers for 3-class (neutral, stress and amusement) classification. While the SVM classifier achieved the worst result with an accuracy rate of 59.56%, the RFC classifier achieved the best result with an accuracy rate of 67.56%. Schmidt et al. [105] used only ECG data and compared many classifiers. However, in the three-class classification, LDA has achieved a higher accuracy rate of 66.29%. Lin et al. [107] utilised the MMSF model to classify three emotions. They also investigated the effect of different types of data on the accuracy rates. The results showed that respiration, EMG and ECG signals were the highly influential signals based on the classification accuracy rates. The researchers achieved a classification accuracy of 83% on three

classes by using all chest signals. However, the accuracy rate decreased to 78% when only ECG signals were used.

When we examine CNN models, Sarkar et al. [108] used Amigos, Dreamer, Swell and WeSAD ECG datasets to train the single-task CNN model. They achieved 95% accuracy rates when they used the WeSAD dataset in testing to classify 4 emotions: baseline, stress, meditation, and amusement. Tanwar et al. [110] used ECG, EMG, body temperatures, EDA and respiration data to train the CNN-LSTM model for 3-class classification. While they used all of the data of 14 people in training, they used the data of 1 person in the testing. Under these conditions, they were able to classify baseline emotion with a 96.63% accuracy rate, stress with 90.20% and amusement emotional status with 72.12%. Behinaein et al. [109] trained the CNN model with WeSAD and Swell datasets and used a Fine-tuned (FT) Leave-one-subject-out (LOSO) validation model. While an 80.4% accuracy rate was obtained without FT, a 91.1% accuracy rate was obtained when FT was applied to 10% of the test data. The parameters of FT must be tuned very well. Using FT may extend the training period. In addition, there is a possibility that the model will get stuck in local minima, and using FT is a task-oriented solution and may not be compatible with every model.

Granero et al. [111] extracted several features from electroencephalography (EEG), ECG, Galvanic Skin Response (GSR) and respiration data. They tested 56 features on ECG data and they observed more representative features according to their emotion classification performances using 14 classifiers. RF, RF with Multi-Class Classifier (MCC) and Bagging classifiers achieved the highest classification accuracy as 75.46% for positive, 75.46% for neutral and 88.34% for negative emotional statuses. Some of the time domain features (minimum, maximum, median, standard deviation RR interval etc.) they used were also used in our study because they achieved successful results in emotional status classification. In Table 2.4, there is a summary of state-of-the-art studies in emotional state classification.

Template matching algorithms calculate the distance between two templates using different metrics. Since each metric obtains a different distance or similarity score,

TABLE 2.4: State-of-the-art studies in emotional state classification

Authors	Datasets	# of Subjects	Recordings	Methods	Performances
Bras et al. (2018) [70]	Emotional dataset	25	Total 25-min per emotion 420-sec training per emotion remaining data for testing neutral, disgusted, and fearful	Kolmogorov complexity theory with Finite-context models and normalized relative compression KNN (k=1) classifier	Accuracy rates in biometric identification: 98.5% for Neutral, 97.2% for Disgust and 97.2% for Fear Accuracy rates in emotion classification: 77% for fear, 74% for neutral and 62% for disgust
Li et al. (2023) [102]	Dreamer Decaf ECG-ID HCILAB	23 30 90 10	Several subjects and Training/Testing ratios	6 distance metrics 6 DL models Set-Group Distance Measure	Emotion classification accuracy rates for HCILAB Dataset: Set-Group Distance Measure: 90% Collaborative-Set Measurement: 10% DNN: 74.07%, 1D-CNN: 77.88% Bi-LSTM: 70.79%, GRU: 73.57%
Vaz et al. (2023) [104]	WeSAD	15	EDA,EMG and ECG Total: 19-min was used 15 segments= 5-min time length and a 4-min overlap 109 features were computed for each segment Several data balancing methods	55 features selected LDA and RFC classifier 80% training / 20% validation, 7 classifiers	3-class Activity classification accuracy rates: LDA: 62.7%, RFC:92% 2-class Activity classification accuracy rates: RFC: 92%
Garg et al. (2021) [106]	WeSAD	15	EMG, EDA, body temperatures, respiration and ECG Stress, neutral, amusement	10-sec sliding windows standard deviation, mean, minimum and maximum values features LOSO cross validation for training and testing separation RFC, SVM, KNN, LDA and AdaBoost	3-class activity classification accuracy rates: SVM: 59.56%, AdaBoost: 64.34% KNN: 65%, LDA: 67.06%, RFC: 67.56% 2-class activity classification accuracy rates: SVM: 76.01%, AdaBoost: 82.24% KNN: 77.26%, LDA: 78.47%, RFC: 84.17%
Lin et al. (2019) [107]	WeSAD	15	Sliding windows (1-sec window with 0.25-sec shift) Blood volume pressure, ECG, accelerometer EDA, temperature Stress, neutral, amusement	MMSF model DL with LSTM, RFC, SVM Logistic regression classifiers 7 subjects in the test set LOSO cross-validation on data from 7 folds	Classification accuracy rates: MMSF: 83%, Logistic regression: 82% SVM: 83% RFC: 85% Only ECG: 73%
Sarkar et al. (2020) [108]	Amigos Dreamer Swell WeSAD	40 23 25 17	Varied recording time 90% training / 10% testing Various activities	10-fold cross-validation 6 signal transformations	Mean classification accuracy rates: Amigos: 79.6% arousal- 78.3% valance Dreamer: 77.1% arousal- 74.9% valance WeSAD: 95% Swell: 92.6% arousal- 93.8% valance - 90.2% stress
Tanwar et al. (2022) [110]	WeSAD	15	ECG, EMG, EDA, temperature and respiration 5-sec time window with 2-sec shift	14 subjects' data for training, 1 subject's data for testing CNN-LSTM	Emotion classification accuracy rates: Baseline: CNN: 96.63%, Stress: 90.20%, Amusement: 72.12%
Behinaein et al. (2021) [109]	WeSAD Swell	15 25	30-sec time window with 1-sec time shift	1%, 5% and 10% FT LOSO validation CNN model	Emotion classification accuracy rate: WeSAD: Without FT: 80.4%, 10% FT: 91.1% Swell: Without FT: 58.1%, 10% FT: 71.6%
Granero et al. (2016) [111]	Private	47	EEG, ECG, GSR and respiration 30-min data Mindfulness, watching ads	56 features 14 classifiers MCC, Bagging and MultiClass classifiers for ECG	Emotion classification accuracy rate for HRV: Positive: 75.46% Neutral: 75.46% Negative: 88.34%
Ours (Phases 2 and 3) (ML) [2]	WeSAD (baseline, stress and amusement)	15	5-min per activity 80% Training/20% Testing, 60% Training/50% Testing and 50% Training/20% Testing sample cases	4-s time windows Manhattan and Euclidean distance features KNN,SVM, DT, BT, S-KNN and NN NB and DT for biometric verification	The best activity classification accuracy rates: S-KNN RespiBAN: 93.67% - 80.29% Min/Max EERs: RespiBAN: 8.32%- 15.43%

it has been frequently compared in the literature [82, 93, 112].

Lee et al. [112] compared the data, in which they segmented 1 heartbeat, 2 heartbeats and 3 heartbeats, in the biometric identification model with different template matching methods. They compared Euclidean distance, Manhattan distance,

cosine similarity and cross-correlation metrics with ECG data collected from the wrist and the finger. ECG data measured from the wrist have a higher accuracy rate with the Euclidean distance method than others. However, when 1 heartbeat segmentation was performed, the Manhattan distance method was observed to perform better in the ECG data collected from the finger. Perpinan et al. [113] explored the k-means algorithm's performance on metabolic syndrome classification using heart rate variability features with Manhattan and Euclidean distance metrics. They found that the Manhattan outperformed the Euclidean distance metric [2]. Rahiem et al. [82] compared 11 distance metrics with KNN (k=19) and SVM classifiers in their biometric authentication model and they stated that the standardized Euclidean distance metric had higher estimation rates than others. Kim et al. [93] compared the EER results of 7 distance metrics in the biometric verification model. The Euclidean distance achieved the lowest EER result, followed by the Manhattan distance. The most unsuccessful one was the Chebychev metric.

2.6 Open Challenges

Studies on ECG have generally focused on disease recognition. Later, it began to be used in biometric systems and it was observed that ECG was affected by environmental and biological factors. As seen in the literature, many studies have emphasised that ECG is affected by physical activities, emotional states and even lifestyle (eating and drinking patterns, alcohol or drug use, etc.). Although it was recognised that these effects had a negative impact on the performance of biometric models, no action was taken to address this issue.

The training process is very important in biometric verification models. Factors such as the number, quality, diversity of data, general recording time, etc. affect the training process. A well-trained model should be generalisable to various data and should not have problems such as over-fitting or under-fitting. For that reason, the number of subjects and devices, and the circumstances in which the

data was gathered are important in terms of investigating the generalisability of the biometric model.

Authentication time is directly related to sampling time. Considering that the number of samples used is kept constant, the shorter the time used for enrollment (data is collected from the subject for enrollment in a shorter time), the faster biometric verification occurs. In biometric models, longer enrollment time generally results in lower error rates. Thus, models that will have low error rates during short enrollment and authentication times should be investigated.

Ease of use is another thing to consider. When it comes to ECG data, it is not desirable to collect data for a long time with medical devices, various cables, disturbing sensors, etc. For this reason, wearable technologies such as smartwatches or T-shirts need to be used, and this creates a more realistic scenario. However, the accuracy of the data provided by wearable devices (providing a noisier signal than a medical device, the distance of the measuring device to the heart, etc.) still needs to be investigated. In addition, the fact that the biometric model produces results quickly provides ease of use.

Although many datasets have been used for model comparisons in the literature, it has not been possible to make a direct device comparison due to the lack of datasets obtained simultaneously from the same participants using different devices. Our study also investigated direct device performances in terms of biometric verification models.

The aforementioned open challenges need to be investigated in the field of ECG biometrics. These challenges are the impact of external factors, including physical activities and emotional states, on ECG-based biometric verification. Additionally, the effects of training data, enrollment and verification times, and the performance evaluation of different devices using ML and DL models and their parameters remain critical to obtain trustworthy biometric verification model. Although it is seen that some of these parameters are frequently investigated and compared in the literature, there is no comprehensive comparison including different device types for biometric verification models.

2.7 Proposed Framework

The open-to-research topics mentioned in Sections 2.5 and 2.6 formed the basis of this thesis. As a first step, the biometric verification performances of medical and wearable devices were investigated with the traditional ML and DL structures. In this study, the effects of parameters such as various classifiers, recording period length, enrollment times, sampling period, quantity of subjects, and traditional and newly produced features on the model were investigated. This step is specified as *Phase 1* in the thesis.

As a second step, a new framework was created to increase the stability of ECG biometric models. Within this framework, physical activities and emotional states affecting the model were examined separately. In this step, activity and emotional state classification was made. The parameters compared throughout the classification are newly produced features, sizes of training and testing sample sets, classical ML classifiers, different DL models and the effect of different image types, number of epochs and optimiser types on these DL models. This step is defined as *Phase 2* in the thesis.

The final step involves examining the data classified according to their activities and emotional states in the biometric verification model separately for each emotion or activity class. Since the data in each class show similar characteristics to each other, its stability has increased and it has been observed that the error rates of biometric verification models have decreased. One of the important points here is that the data is grouped according to the assigned/classified activity labels, not the original activity labels. Considering the real-life scenario, when the model assigns the data of a person actually walking to the running class, ignoring the error here will reduce the reliability of the system. There is no study in the literature that takes this situation into consideration. In addition, all parameters explained in *Phase 1* and *Phase 2* were also examined at this stage (i.e. *Phase 3*).

The general methodology and detailed examination of each phase will be explained in the following Chapters.

Chapter 3

Methodology of the ECG Biometric Studies

In this chapter, general information about methodology in terms of selected sample datasets, pre-processing, extracted features, image representations, classifiers and biometric performance assessments are explained.

3.1 Datasets

The monitoring of the electrical activities of the heart has a long history within medical and research domains. Although the first Electrocardiogram (ECG) recorder was found in 1906 by Nobel Prize winner Willem Einthoven [114], the recording of the heartbeat of mammals first started with animal experiments, and then it was followed by application to humans [115, 116]. In this context, the portable Holter ECG monitoring device was developed by Norman Holter in 1983 [117]. Holter devices can monitor ECG recordings for periods between 24 hours and 72 hours and therefore can be used for the cardiac arrhythmia detection [116]. With the release of many ECG recorders, the format, the sample recording times and technical features of the devices have changed resulting in implementations of chest straps [118], arm bands [119] and smart watches [120] etc. For this reason, many

datasets were created using different ECG recorders. These datasets can be either public or private due to data protection laws and ethical considerations. A great number of ECG diagnostic methods have been proposed using ECG datasets using data from healthy, quasi-healthy (i.e. no significant arrhythmia but containing signal corruptions) and unhealthy (having heart disease/diseases) participants. The most common datasets were listed in Table 3.1.

TABLE 3.1: Commonly used public and private ECG datasets

Datasets	Name	Recording Lengths	# of Leads	# of Subjects
Private	BigIdeasLab_STEP [121]	1 hour	1	53
	Stanford [122]	30 seconds	1	53549
	AHA [123]	3 hours	2	155
	Mayo CV [124]	10 seconds	12	180922
	CCDD [125, 126]	10 seconds	12	193690
	THEW [127]	24 hours	3	201
	CEBSDB-I [128]	1 hour	2	21
	CEBSDB-II [129]	1 hour	3	17
	WECG [130]	30 minutes	1	22
	UofTDB [75]	2-5 minutes	1	1012
	CSE [131, 132]	10 seconds	15	125
Public	PTB Diagnostic ECG [133]	10-30 minutes	15	290
	European ST-T [134]	2 hours	2	79
	AF Classification Challenge 2017 [135]	9-61 seconds	1	12186
	PTB-XL [136]	10 seconds	12	18885
	ECG-ID [137]	20 seconds	1	90
	WeSAD [105]	2 hours	3	15
	E-HOL-03-0202-003 (E-HOL) [138]	24 hours	3	202
	Vollmer [3, 139]	20 minutes	3 and 1	13
	MIT-BIH Arrhythmia [140]	30 minutes	2	47
	MIT-BIH Noise Stress Test [141]	30 minutes	12	15
	Ventricular Tachyarrhythmia [142]	8 minutes	12	35
	MIT-BIH Malignant Ventricular Arrhythmia [143]	30 minutes	12	22
	MHealth [144]	15 minutes	2	10
	St Petersburg [145, 146]	30 minutes	12	32
	Shaoxing Hospital [147]	10 seconds	12	10646
	Fantasia [148]	120 minutes	12	40
	LT-AF [149]	24 hours	12	44
	CYBHI [150]	60 seconds	2	128
	STAFF III [151]	10 minutes	12	35
	ICBEB [152]	6-60 seconds	12	6877

Electrodes, which are conductive pads affixed to the skin, facilitate the capture of electrical currents. These electrodes detect the heart's electrical activity. Each

ECG lead corresponds to a graphical representation of the heart's electrical behavior. The computation of an ECG lead involves analysing the electrical currents recorded by multiple electrodes [153]. Essentially, an ECG lead reflects the comparison of electrical currents at various measurement points. Medical devices usually have three or more leads. Wearable devices usually have between 1 and 3 leads. Obtaining ECG data from various parts of the body through many leads is important for disease diagnosis and improving signal quality [153].

The Stanford [122], Mayo CV [124], AHA [123] and CCDD [125] datasets contain ECG recordings of thousands of people taken in the hospital environment to diagnose heart diseases. Especially the CCDD dataset [125] is updated every year and new data has been added for more than 10 years.

Consumer-based wearable devices and lab-based wearable sensors can measure ECG and Heart Rate (HR) signals. For instance, WeSAD [105], WECG [130] and BigIdeasLab [121] datasets were collected from wearable devices. Scientists have used data from wearable sensors and compared the results of wearable ECG recorders and classical 12-lead ECG recorders data. The UofTDB dataset was collected on a single day during sitting, standing, exercising, supine and squat/tripod activities from fingers [75].

There are many private datasets because researchers collect their data for just their studies. Moreover, they can combine these collected ECG recordings with available datasets. To be eligible for use in our research, public datasets must meet to specific criteria. These criteria include the inclusion of data from only healthy subjects, coverage of various physical activities and emotional states, and a recording duration that is sufficiently long to prevent over-fitting. Furthermore, the availability of diverse types of devices that can record simultaneously aligns with the primary objectives of our study. The public E-HOL (long recording time, large number of subjects), WeSAD (several emotional states from wearable device) and Vollmer (different physical activities from various simultaneous ECG recording devices) datasets that overlap these criteria were found and used in our study. In the following subsections 3.1.1, 3.1.2 and 3.1.3, the datasets which

were used in these studies are explained in detail. The E-HOL [138] dataset was selected because it provides a large sample of healthy subject data from medical Holter device recordings with a variety of physical activities and emotional states. The WeSAD [105] was selected because it provides wearable chest band ECG recordings from different emotional states and the Vollmer et al. dataset [139] includes ECG data which were collected on different physical activities from 5 medical and wearable-based devices. This dataset is quite useful for comparing device performances. In addition, the lack of studies in the literature on biometric verification performance evaluations using these datasets played a predominant role in dataset selection.

3.1.1 E-HOL-03-0202-003 (E-HOL)

24 hours ECG recordings from 202 healthy subjects were recorded by the SpaceLab-Burdick digital Holter device in the E-HOL [127, 138] dataset. The device has 3 electrodes which were located according to the pseudo-orthogonal lead configuration. Participants in the dataset have no cardiovascular disease and condition, high blood pressure problem, chronic disease, medications and pregnancy. The population of subjects consists of 100 females, 100 males and 2 undefined people. The data were collected at 200 Hz sampling frequency with 10 uV amplitude resolution after a 20-minute resting/supine period.

The Holter device is an ambulatory device. For this reason, there is no restriction for activities [154]. Being asleep and awake were stated, but other activities were not specified in detail.

3.1.2 Multimodal Dataset for Wearable Stress and Affect Detection (WeSAD)

The WeSAD dataset [155] includes synchronised data from a chest-worn RespiBAN Professional device [156] and a wrist-worn Empatica E4 device [157]. These devices

can measure ECG, Electromyography (EMG), respiration, temperature, Electrodermal activity (EDA) and acceleration data. 17 healthy subjects donated samples for data collection, however, the data of 2 subjects had to be discarded by Schmidt et al. [105] because of sensor malfunction. Therefore, the dataset includes the data of 15 subjects (3 females and 12 males) [105, 155]. The RespiBAN device can measure ECG signals via standard three-point ECG sensors around the chest with a 16-bit analogue-to-digital converter resolution at 700 Hz sampling frequency. The dataset provides single channel ECG signal.

The dataset consists of neutral/baseline, stress and amusement emotional state ECG recordings. In the baseline condition, subjects can sit or stand at a table and read magazines for 20 minutes whilst wearing the device. In the amusement condition, a selected video was watched by subjects for approximately 6 minutes [158]. Subjects have a public speaking task and mental arithmetic tasks for 10 minutes in the stress condition. There are 7 minutes of meditation time in a sitting position between amusement and stress conditions to return the subject to neutral emotional states.

Handling unbalanced datasets in classification tasks is crucial for achieving accurate and reliable results. When the distribution of samples across different classes is imbalanced, it can lead to biased model performance [159]. Some techniques to handle unbalanced datasets include using different evaluation metrics, resampling and class weights, etc. Each technique can increase the computational cost, complexity and risk of over-fitting [160, 161]. Additionally, the results of these techniques vary for each dataset and situation. In our case, if the amount of data obtained from each activity is not equal, one activity may not be learned at the same rate while another activity is learned more during training. This makes the classification process difficult and affects the prediction rates. In addition, using unbalanced data affects the number of samples in Phase 3 and the framework's reliability. We propose a biometric verification framework that can be used for different datasets and devices, and we test the same experimental procedure on different datasets. For this reason, instead of using different techniques, we used

the 6 minutes of data from each activity (18 minutes total) in the WeSAD dataset in all experiments.

3.1.3 Simultaneous physiological measurements with five devices at different cognitive and physical loads (Vollmer)

Vollmer et al.[3] conducted experiments with ECG signals from 13 healthy participants which were collected from five different ECG recorders started simultaneously to create a dataset [139]. These devices can be grouped as medical-based (i.e. clinically certified) such as NeXus-10 MKII [162], eMotion Faros 360° [163], SOMNOtouch NIBP [164] and consumer-based products such as Hexoskin Hx1 [165] and Polar RS800 Multi [18]. These devices can measure ECG, heart rate variability, heart rate, saturation, photoplethysmogram and pulses. The Vollmer dataset includes ECG signals during 4 tasks: resting, treadmill walking at a speed of 1.2 m/s, standing still position and uphill walking on the treadmill with 15% track inclination at the same speed. Each activity takes 5 minutes. Information about each device and the sensor placement are explained in Table 3.2. Attachable patches are standard Ag /Ag-Cl electrodes while the chest strap has flat electrodes.

TABLE 3.2: Device specification of Vollmer dataset

Device Name	# of Sensors	Locations of Sensors	# of ECG channels	Sampling Frequency	ECG Electrode Type
SomnoTouch	6	Left chest, Left & right abdomens	3	512 Hz	Attachable patches
Emotion Faros 360	3	Left right chest, right abdomen	3	1000 Hz	Attachable patches
Nexus-10 MKII	4	Left right chest, right abdomen	3	8000 Hz	Attachable patches
Hexoskin Hx1	4	Top left right abdomens, bottom left abdomens	3	256 Hz	Textile
Polar RS800 Multi	1	Top centre abdomen	1	1000 Hz	Bipolar Chest strap

Except for the Polar device, the other devices are able to record raw ECG signals. However, the Polar device can measure R-peaks (seen in Fig.1.3) which were used as a reference point for the other devices' measured R-peak position synchronisation [3]. All devices have different sampling frequencies in this dataset. The Polar device R-peak locations were used by Vollmer to equalise all devices to the same sampling frequency (the lowest frequency is 256 Hz) and all heartbeat locations [3]. The dataset provides both raw ECG signals and synchronised signals in Physionet [146]. The provided synchronised signals were used in our experiments. This is because ECG signals at different sampling frequencies and having a shift or difference between the locations of the R peaks despite measuring simultaneously may affect the activity classification and biometric verification performances.

3.2 Data Analysis

3.2.1 Pre-processing

Pre-processing is vital in terms of increasing the usability of ECG data and the performance of the system. All 3 datasets used in our studies have different devices, different sampling frequencies, different data acquisition processes, different sensor types and different usage purposes. These parameters are crucial to understand the data structure and possible noises into the signals. For this reason, two different data pre-processing methods, which were listed below, were used in our experiments.

The sampling frequencies of all datasets (700 Hz in WeSAD and 256 Hz in Vollmer datasets) was reduced to 200 Hz. 200 Hz was selected because the medical Holter device in the E-HOL dataset has this frequency and finding the fiducial point of ECG at this frequency is the more common analysis reference point. Since the WeSAD dataset has a single-channel ECG signal, a single channel was used in the other datasets (X was selected in the XYZ configuration). In addition, using

single-channel data allows easy comparison of different datasets, and especially data from wearable devices.

1. Since only ECG records of healthy subjects taken from medical and wearable devices were used in our study, a low-frequency noise removal filter and a powerline band-removal filter were used on the EHOL and the WeSAD datasets as the first pre-processing method [1]. Specifically, the high-pass Butterworth filter with 0.5 Hz cutoff frequency and a notch filter with 50/60 Hz (depending on the power source) have been used for baseline wander, zero drift and powerline frequency removal. These filters have been selected considering previous studies with E-HOL [154, 166] and WeSAD [167, 168] datasets. This pre-processing method was implemented using Python *neurokit2* library [168].

2. The Vollmer and WeSAD datasets were filtered using a second pre-processing method. Both datasets include wearable devices and attachable patches to record ECG signals. For this reason, the noise produced by muscles and friction connected to activities may be introduced to the signal. The signals provided from the Vollmer dataset have already been pre-processed using a trimmed moving average filter, and Z-score normalization [3]. However, it has been observed that these filters are insufficient to denoise the signal and it is shown in Fig.3.2. A band-pass filter was used to obtain signals in a similar range as the first preprocessing method. Specifically, a three-order Butterworth band-pass filter with 0.5 Hz low cut-off frequency and 45 Hz high cut-off frequency and a mean filter were used to remove noise and provide signal smoothing [2].

Due to the different structure of the datasets, two different preprocessing methods were used. Since the Holter device is a medical device, it can contain a lot of sensitive information and it has the potential to have higher accuracy. For this reason, the first filter system (cf. 1) was designed to eliminate the most prominent noise in the E-HOL dataset, but a sharp band gap was not used to avoid losing sensitive information. On the other hand, the Vollmer dataset has been previously

filtered and the sampling frequency has been changed [3]. For this reason, the second filter system similar to the previous system was used to reduce both the number of filters and the number of processes. In this system, instead of using three filters (1 high-pass and 2 notch filters), a Butterworth band-pass filter is used. Using a band-pass filter with 0.5 Hz and 45 Hz cut-off frequencies saved us from using 2 notch filters and gave very similar signal results.

3.2.2 Extracted Features for Machine Learning Models

Classical machine learning models need features for classification. As explained in detail in the Chapter 1, each heartbeat consists of 5 important segments which are named as P, Q, R, S and T peaks. When analysing ECG data, there are 2 different approaches; fiducial and non-fiducial [169]. The fiducial feature extraction approach that is used in this study is related to P, Q, R, S and T peaks, their amplitudes, and time intervals between each wave [1, 2]. Along with the fiducial features that are used by the majority of the studies in the literature, new features derived using P, Q, R, S, and T waves are explained in 3.2.2.1 and 3.2.2.2 sections.

3.2.2.1 Fiducial and time interval features

It is desirable to determine the information of each ECG signal fiducial point in terms of their uniqueness and therefore their importance as a feature for biometric verification. To extract features, E-HOL and WeSAD data were filtered and cleaned using the first pre-processing method.

After the pre-processing, N samples per recording from each subject were cut from the ECG signals. Each sample represented a time window with size T and a delay between windows of s. Its dimension can be shown as $[(N\_subjects \cdot N) \times (T + 1)]$. Each sample was formed with a 10-second window size and a 1-second delay between windows was used for feature extraction. The final data structure after feature extraction can be represented as $[(N\_subjects \cdot N) \times (N\_features + 1)]$.

15 features are selected from 9 fiducial features used in the literature and from the intervals between heartbeats [5]. The features we used have been used in various combinations in the literature and their effectiveness has been investigated [5, 170, 171]. However, in our study, it was used for biometric verification for the first time in E-HOL and WeSAD datasets. In the feature extraction process, *P, Q, R, S and T* peaks of the signals were found as a first step using the *neurokit2* library inspired by the Pan-Tompkins algorithm [168, 172]. The position of R peaks, which was a reference to find other peaks, was found by the *ecg_peaks* method implemented in the Python *neurokit2* toolbox [11]. The maximum P and T values, or the minimum Q and S values, were determined within 200 milliseconds (ms) for P peaks, 40 ms for Q peaks, 40 ms for S peaks and 340 ms for T peaks interval before or after the reference R peak [172–174].

Finding fiducial points is important for calculating other features. The detected P, Q, R, S and T peaks are shown with different markers in Fig. 3.1 on the ECG signals obtained from the subjects in the different datasets. Each time window acquired throughout the ECG signal begins with the R peak. However, the endpoint of the time window (in other words, the end point of the ECG signal) varies for each person due to physiological (age, weight, etc.) and environmental factors. Some signals terminate after a complete PQRST cycle, while in others part of the cycle is contained within the time window (as in Fig. 3.1). In order to determine the starting and ending points of the signals in each window and to obtain consistent time windows, the last heartbeat with all peaks detected are not marked. Moreover, the first heartbeat is not marked because it starts with the R peak and does not include a complete cycle.

QS, PQ and ST which represent distances between peaks are found using the maximum value positions of two peaks. The mean of the distances between the two peaks obtained for each recording was calculated. P, Q, R, S and T are *peak amplitudes* which are calculated using the mean of the maximum value within the recordings. Extracted fiducial features and R-R interval features are shown in the Table 3.3.

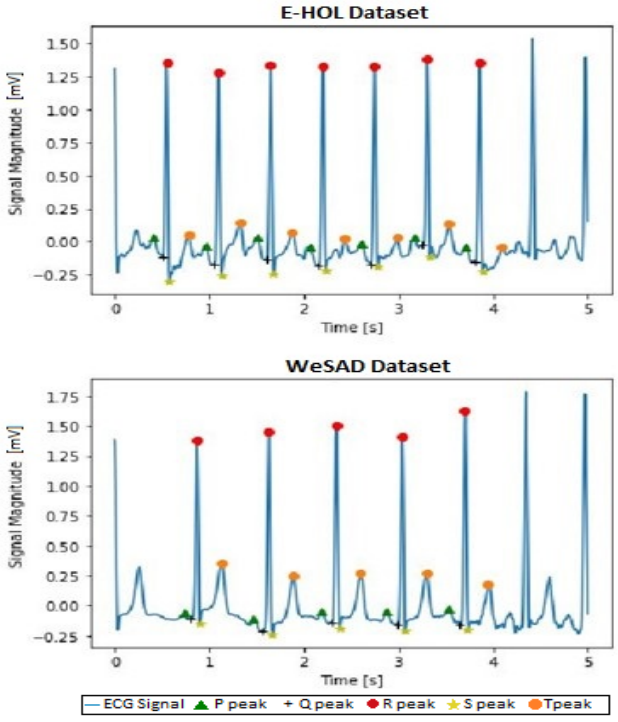

FIGURE 3.1: P, Q, R, S and T peaks detection for different subjects in E-HOL and WeSAD datasets

Prakash et al. [175] formulated the Heart Rate (HR) as *HR=60/RR interval in seconds* Beat-per-minute (Bpm). Although the minimum difference between two consecutive R peaks is assumed to be 0.2 seconds (i.e. HR=300) [176], the maximum HR of healthy subjects during exercise is considered to be 220 bpm [177]. The sampling frequency of WeSAD and E-HOL datasets is 200 Hz. If the distance between the two R peaks is shorter than 50 points, the HR will be higher than 240 bpm. It also indicates unhealthy subjects (having tachycardia) or incorrect detection of R peaks. All ECG data were declared as healthy subjects. Therefore, *RR50p* and *RR50pRatio* were calculated to check whether the locations of the R peaks were correctly detected. *RR50p* is the number of occurrences of R to R distances shorter than 50 points while *RR50pRatio* checks the occurrences of all samples. All *RR50p* and *RR50pRatio* values were found as zero. In this way, it was validated that the R peak locations were determined correctly.

In this study, we used the data obtained from wearable and medical devices with temporal and R-R statistics features that are frequently used in the literature

TABLE 3.3: Details of the extracted features. t is the time vector of the sample and N is the number of heartbeats in a sample. P, Q, R, S and T are wave peaks. [4, 5]

Features	***Description***	
QS	$\frac{1}{N}\sum_{i=1}^{N} t_{S_i} - t_{Q_i}$	Distances between peaks (mean value for each sample)
PQ	$\frac{1}{N}\sum_{i=1}^{N} t_{Q_i} - t_{P_i}$	
ST	$\frac{1}{N}\sum_{i=1}^{N} t_{T_i} - t_{S_i}$	
P_{amp}	$\mu\{P_i\}_{i=1}^{N}$	Peak amplitude (mean value for each sample)
Q_{amp}	$\mu\{Q_i\}_{i=1}^{N}$	
R_{amp}	$\mu\{R_i\}_{i=1}^{N}$	
S_{amp}	$\mu\{S_i\}_{i=1}^{N}$	
T_{amp}	$\mu\{T_i\}_{i=1}^{N}$	
$minRR$	$\min\{t_{R_i} - t_{R_{i-1}}\}_{i=2}^{N}$	R-R interval features
$maxRR$	$\max\{t_{R_i} - t_{R_{i-1}}\}_{i=2}^{N}$	
$medRR$	$m\{t_{R_i} - t_{R_{i-1}}\}_{i=2}^{N}$	
$meanRR$	$\mu\{t_{R_i} - t_{R_{i-1}}\}_{i=2}^{N}$	
$stdRR$	$\sigma\{t_{R_i} - t_{R_{i-1}}\}_{i=2}^{N}$	
$RR50p$	$\dim(A), A = \{RR50p \mid RR50p = RR_i < 50\}$	
$RR50pRatio$	$\frac{RR50p}{\dim(RR)}$	

because the aim of this study is to determine the baseline of EER for different types of devices and subjects in biometric verification models.

In the classical machine learning model, the distinctiveness of the features in the training and testing sets is important for the performance of the model. In addition to the features in this section that allow us to determine baseline EERs and comprehend device performances, we also investigated the performance of novel features in ECG biometric verification models.

3.2.2.2 Manhattan and Euclidean Distances as features

The performance evaluations of newly discovered features were utilised for activity/emotion classification and biometric verification since we used datasets containing physical activities (Vollmer) and emotional states (WeSAD).

Window size selection within a time-based sample is another key effect on the biometric system performances. Abdul-Kadir et al. [178] sought an optimum window size for ECG atrial fibrillation (AF) recognition. They found that 4 seconds are the best option for their features with NN and SVM classifiers (95.3% and 95% AF recognition accuracy rates respectively). Inspired by the study, we investigated the effects of different time windows on the biometric verification model. For this reason, we used only 4 seconds time windows on this study. Each sample was defined with a 4 second window size and a 1 second delay between windows. To make a clear separation without losing any Q, R and S peaks and prevent over-fitting, a 1 second delay between windows was selected. All time windows started with the heartbeat with all peaks detected.

28 fiducial features were extracted from each sample in the pre-processed matrix. As utilised a previous feature extraction method, Q, R, S and T peaks were determined in the first step. Sedghamiz et al. [179] developed the *BioSigKit* toolbox to find Q, R, S and T peaks detection. Initially, this toolkit was inspired by Pan Tompkins Q, R and S detection implementation [172]. MatLab R2020a was used for the feature extraction and the second pre-processing method. The pre-filtered ECG signals provided by Vollmer [3] appears to be insufficient in eliminating noise. For this reason, the second filter method (cf. 3.2.1) is used and a smooth signal is obtained. A sample of Q, R, S and T detection is represented in Fig.3.2.

Finding the P and T peaks is a challenging and often uncertain task, especially in the activity case. For this reason, Q, R and S peaks were selected.

The Manhattan (also known as taxicab geometry, City Block or L1) distance is defined as the distance between two points in a uniform grid where you can only move horizontally or vertically. In a city map that has a grid-like structure, the Manhattan distance between two points is defined as the shortest path between them that consists of horizontal and vertical segments only whilst Euclidean distance is the closest and straight line path between two points. The general formula of the Manhattan distance and Euclidean distance are shown in Eq.(3.2) and Eq.(3.1), while $d(x,y)$ or D is the distance between x and y points, n represents the number

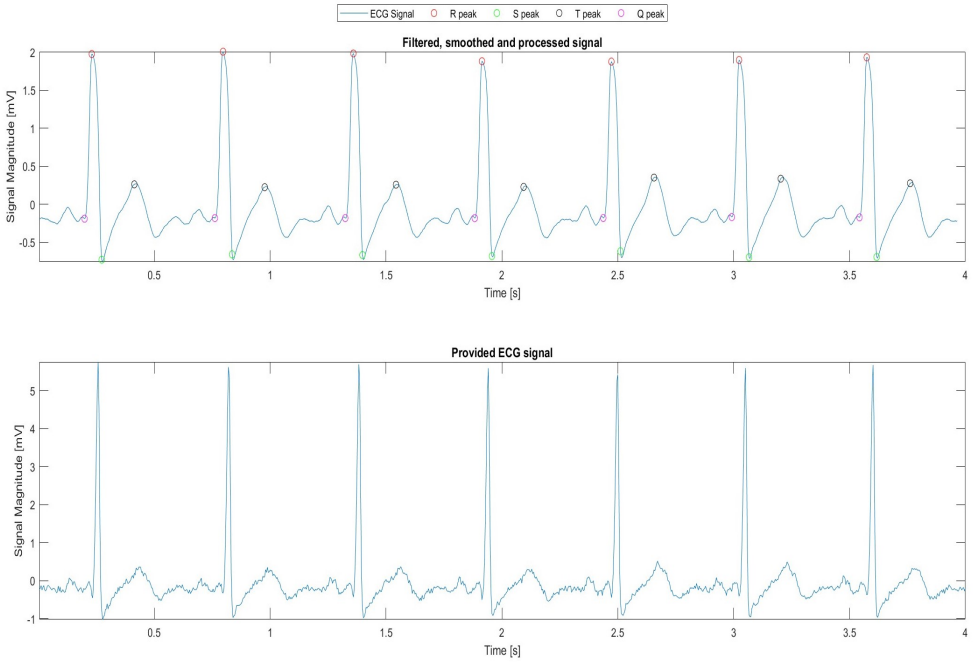

FIGURE 3.2: The figure represents provided pre-filtered ECG signals [3] and processed signals from a Hexoskin wearable device for 4 seconds. Q, R, S and T peaks are shown in different colours.

of dimensions, x represents x-coordinates, y represents y-coordinates and i represents the i_{th} number of dimensions. The p parameter in Eq.(3.3) is originally used for other distance function calculations. For the Euclidean distance, *p=2* while *p=1* for a Manhattan distance calculation.

$$\text{Euclidean Distance} = d(x, y) = (\sum_{i=1}^{n} |x_i - y_i|^2)^{\frac{1}{2}} \tag{3.1}$$

$$\text{Manhattan Distance} = d(x, y) = (\sum_{i=1}^{n} |x_i - y_i|) \tag{3.2}$$

$$\text{Minkowski Distance} = d(x, y) = (\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}} \tag{3.3}$$

$$\text{Hamming Distance} = D_H = (\sum_{i=1}^{n} |x_i - y_i|)$$
$$x = y, \quad D = 0$$
$$x \neq y, \quad D \neq 1 \tag{3.4}$$

Eq.(3.4) is often used to see if two templates match. It is also useful for Boolean and string vectors [180]. The main differences between Euclidean and Manhattan distances are represented in Fig.3.3 [181].

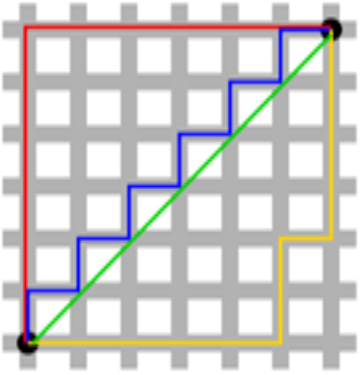

FIGURE 3.3: The green line represents the Euclidean distance. The blue, red, and yellow lines, all of which have equal units, represent the Manhattan distance.

It is clearly seen that Manhattan distance paths (red, yellow or blue) are equal to 12 units while Euclidean distance has a shorter length of approximately 8.49 units. The Manhattan distance value is usually higher than the Euclidean distance value because the Manhattan distance is based on the realistic path between two points. Euclidean distance generally performs worse than Manhattan distance in high-dimensional (more than 3D) data and unnormalized data [182]. Our data is not high-dimensional. However, Manhattan distance gives better results in discrete and binary data [182]. These two metrics are investigated because of the situations where both distance metrics are advantageous.

The same-peak, 2-peak and 3-peak combinations of Manhattan distances and Euclidean distances were assessed for activity classification accuracy. To clarify the extracted features, we labelled the maximum peak value of each P, Q, R, S and T

points as X. Manhattan distances between 2 peaks of the same type are shown in Table 3.4. The same-peak can be represented as the $(XX)_{man} = |(x_{i+1}) - (x_i)|$ formula. Euclidean distances between 2 peaks in the same-peak group is $(XX)_{euc} = \sqrt{((x_{i+1}) - (x_i))^2}$ where x is a selected peak from Q, R, S or T and i is the selected peak location. To calculate 2-peak and 3-peak distances, X, Y and Z can be used instead of Q, R and S. When calculating $(XYZ)_{man}$ with a simple example, the formula $|X - Y| + |X - Z| + |Y - Z|$ is used for all Q, R, S and T combinations. The same procedure can be used with the Euclidean distance formula.

The absolute value of subtraction is always positive, and the absolute value of subtraction is calculated in the Manhattan distance formula. The square of the subtraction parentheses is also always positive, and the square of the subtraction parenthesis is used in the Euclidean distance formula. For this reason, there is no need to calculate the YX distance when calculating the XY distance for both cases. All the features used are detailed in Table 3.4.

TABLE 3.4: Explanations of features are represented as X, Y and Z symbols.

Features	**Values of X,Y and Z**	**Description**	
X	Q_{amp}, R_{amp}, S_{amp}, T_{amp}	$X = \{x_1, x_2, ..., x_n\}$	Peak Amplitudes
XX_{man}	QQ, RR, SS, TT	$d(X_{i+1}, X_i) = \{\lvert X_{i+1} - X_i\rvert\}_n$	Manhattan
XY_{man}	QR, ST, RS, RT, QS, QT	$d(X_i, Y_i) = \{\lvert X_i - Y_i\rvert\}_n$	Distances
XYZ_{man}	QRS, RST	$d(X_i, Y_i, Z_i) = \{\lvert X_i - Y_i\rvert + \lvert X_i - Z_i\rvert + \lvert Y_i - Z_i\rvert\}_n$	
XX_{euc}	QQ, RR, SS, TT	$d(X_{i+1}, X_i) = \{\sqrt{(X_{i+1} - X_i)^2}\}_n$	Euclidean
XY_{euc}	QR, ST, RS, RT, QS, QT	$d(X_i, Y_i) = \{\sqrt{(X_i - Y_i)^2}\}_n$	Distances
XYZ_{euc}	QRS, RST	$d(X_i, Y_i, Z_i) = \{\sqrt{(X_i - Y_i)^2 + (X_i - Z_i)^2 + (Y_i - Z_i)^2}\}_n$	

All possible feature combinations were examined separately according to direct biometric verification performances to find the feature sets that give the highest accuracy rate. According to these analyses, only same-peak-based features have the lowest performance compared to 2-peak and 3-peak features. Even if the general performances of 2-peak and 3-peak were similar, in low training cases, 2-peak features achieved better results than 3-peak features.

The amplitudes of *T peaks* may not be easily found, especially during activity. All analyses included *T-peak amplitudes* had insufficient performances than other peaks in biometric verification. Consequently, extracted features with T peaks were not selected for the final feature vector. Features with the best biometric verification performance were selected as QR_{man}, RS_{man}, QS_{man}, QR_{euc}, RS_{euc}, QS_{euc}.

Temporal features in ECG biometrics change rapidly during physical activity. This may be a distinctive factor, especially in activity classification, because as the intensity of physical activity increases, the time between P, Q, R, S and T peaks will decrease. In addition, the amplitudes of these peaks increase during physical activity, and they are also distinctive for biometric models. Finding the Euclidean distance or Manhattan distance of the time intervals between peaks incurs extra complexity and computational costs. Instead, the distance between the amplitudes of the peaks is a simpler process. Since less computational cost and rapid verification are important factors in biometric verification models, we used the simplest method. The reason why we do not use the new features together with the features in Table 3.3 or other temporal features is to compare the effectiveness of the newly produced Manhattan and Euclidean features.

In this section, two novel features based on Manhattan and Euclidean distances of Q,R and S peaks' amplitudes are introduced to the literature.

3.2.3 Time-Frequency Representations for Deep Learning Models

Deep learning (DL), a subset of machine learning, is a new model compared to classical machine learning. While classical machine learning makes predictions based on extracted features, DL works like a simulation of human brain neurons and there is no need for physical feature extraction. In DL studies, features are calculated by deep networks. In that way, the learning system finds optimal features and uses these features for the learning process.

As seen in the literature, many studies have examined the contribution of different window intervals, the use of 1-dimensional (1D) or 2-dimensional (2D) data, and the use of different CNN structures for DL [78, 79, 88, 183]. Despite the high performance of 1D deep learning structures, it has been observed that 2-dimensional structures provide superior results.

Although there are many existing studies, this study contributed to the literature in terms of 3 different time-frequency representations, examining these representations with different window sizes, using them in activity classification, and examining the effects of activities on biometric verification with different CNN structures.

In this study, ECG data from medical-based Faros and wearable-based Hexoskin devices in the Vollmer dataset were used to create scalogram [184], spectrogram [185] and mel-spectrogram [186] images using 2 seconds, 4 seconds and 10 seconds time windows. These images were examined in activity classification with GoogleNet, ResNet50 and DenseNet201 CNN structures and activity-aware biometric verification with ResNet50 and DenseNet201 CNN structures. The 10 seconds time window is used for ECG studies because it allows for the calculation of the average heart rate over a 10 seconds period [154, 187–189]. The 4 seconds time window has been used in classical machine learning and deep learning studies and has shown that high accuracy rates can be obtained even if the window size is small [178, 190]. The 2 seconds time window is sufficient to classify arrhythmias [191, 192]. In addition, the 2-second time window has been used in human identification and gave successful results [193, 194]. The combinations and performance comparisons of time-frequency representations and pre-configured CNN models for ECG-based biometric identification were first examined by Byeon et al. [83, 195]. They used spectrogram, scalogram, mel-spectrogram, log-spectrogram and MFCC images of ECG signals from patients and healthy subjects in DenseNet, VGGNet, ResNet, Xception and Ensemble CNN models. However, biometric verification was not performed in these studies.

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. Time-frequency representations used in our study are explained in Sections 3.2.3.1, 3.2.3.2 and 3.2.3.3.

3.2.3.1 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 [196]. Spectrograms were created for the visualisation of speech or sounds by Koening et al. [197] and were used for sound recognition and classification purposes [185, 198]. 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.

Short-Time Fourier Transform (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.(3.5) 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 $\mathrm{F}_s^w(t,f)$ is the STFT [199].

$$\mathrm{S}_s^w(t,f) = |\mathrm{F}_s^w(t,f)|^2 \tag{3.5}$$

Each window (2 seconds, 4 seconds or 10 seconds cases) from the pre-processed matrix is used to create a spectrogram image using a *spectrogram* toolbox in MatLab R2022a. Default parameters (no overlapping, Hamming windows) were used to apply the same procedures for all image-frequency representations and to avoid extra computational costs. In the spectrograms, the interval where the signal changes more (e.g. during speech) is shown in brighter colours (yellow), while the interval where the signal changes less (e.g. silence) is shown in darker colours

(blue/navy blue) [200]. A sample of the created spectrogram image is shown in Fig.3.4.

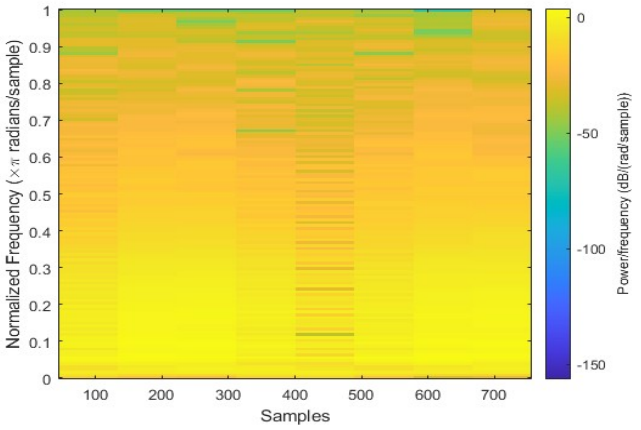

FIGURE 3.4: A sample 4 seconds time windowed spectrogram image of the Subject #1 from the Hexoskin device.

Fig.3.4 is the sample representation of a spectrogram image. However, colorbars, axes and labels were removed from each image feed CNN structures.

3.2.3.2 Mel-spectrogram

A mel-spectrogram is a spectrogram transferred to the mel-scale. According to previous studies, humans perceive frequencies on a non-linear scale [201, 202]. In other words, humans can easily detect differences in low frequencies while they cannot detect differences in high-frequencies. A mel-scale which is widely used in voice analysis accentuates the low-band frequency in voice and eliminates the high-band frequency noise [83]. The mathematical formula of the mel-scale is shown in Eq.(3.2.3.2) [83, 203]. In this equation, f represents the frequency (Hz) and m represents the mel-scale.

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

ECG involves substantial information which is commonly used in ECG applications at the low-band frequency [83, 204]. For this reason, the Mel-spectrogram is investigated with other time-frequency representations in our experiments.

The *melSpectrogram* toolbox [205] in MatLab R2022a was used to obtain the mel-spectrogram. 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 (As a default setting, there are 32 band-pass filters in our experiments.). Each window in the pre-processed matrix was thought of as a channel in this study. The centre frequencies of all filters and the time instants for each analysis window were achieved as outputs. These outputs were used to create a mel-spectrogram for each channel. 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.

Each window (2 seconds, 4 seconds or 10 seconds) from the pre-processed matrix is used to create mel-spectrogram images. Fig.3.5 shows a created mel-spectrogram sample image with a 3D representation and its 90-degree view from above at the 256 Hz sampling frequency. As suggested by the *melSpectrogram* toolbox [205], the original sampling frequency of 256 Hz was used to generate the mel-spectrogram. 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 [206]. 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.

As the given sampling frequency increases, the appearance of the navy blue parts of the image (non-informative parts) changes, but the resolution of windows reduces. In all the images obtained, there are navy blue bands in the same frequency ranges. It has been observed that the lowest frequency without blue bands is 1700 Hz. At this frequency, mel-spectrograms were generated and tested as inputs in some

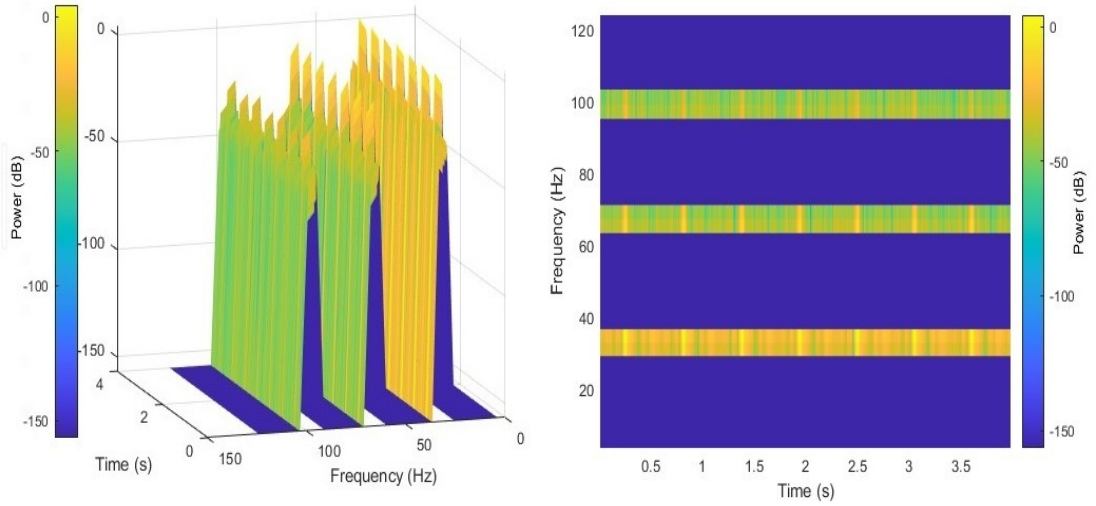

FIGURE 3.5: A sample 4 seconds time windowed Mel-spectrogram image of the Subject #1 from the Hexoskin device. The representation of a 3D view on the left and a 2D view on the right.

activity classification and biometric verification models. It was observed that the execution time in training was prolonged, accuracy rates were lower in activity classification CNN models and EERs were higher in biometric verification CNN models. For this reason, the original sampling frequency of the dataset (256 Hz) is selected to obtain a more precise image and shorter training time. As with the spectrogram images, labels and numbers were deleted from the generated mel-spectrogram graphs and only 2D images presented in CNN structures.

3.2.3.3 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 calculated by taking the absolute value of the signals' CWT. The CWT is achieved by windowing the signal with a wavelet that is scaled and windowed across the sample in time [207]. 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 scalogram is used when

you want to better locate short-duration, high-frequency events in time and low-frequency, longer-duration events in frequency.[207]. The mathematical expression of the CWT is shown in Eq.(3.6) [184].

$$CWT\left(a,b\right)=\frac{1}{\sqrt{a}}\int_{-\infty}^{+\infty}f\left(t\right)*\psi\left(\frac{t-b}{a}\right)dt$$
$$a\in\mathbb{R}^{+}-\left\{0\right\},\quad b\in\mathbb{R} \tag{3.6}$$

In Eq.(3.6), $\psi\left(t\right)$ 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 $\mathbb{R}$ represents *Real Numbers* [184, 208].

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 [209].

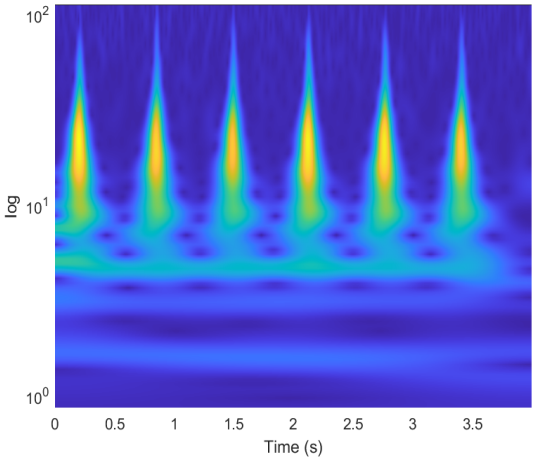

FIGURE 3.6: A sample 4 seconds time windowed scalogram image of Subject #1 from the Hexoskin device.

A sample of a created scalogram image is shown in Fig.3.6. As with other image representations, scalogram images were created with 2 seconds, 4 seconds and 10 seconds time windows without any labels on the images to feed CNN models.

Since ECG signals are sensitive to noise, there are several studies on the frequency domain for noise reduction. If a 1D signal is converted from the frequency domain to the time-frequency domain, the signal becomes 2D. These 2D signals can be

used in CNN models. The spectrogram, mel-spectrogram and scalogram are different methods to transfer signals from the frequency domain to the time-frequency domain. To obtain the frequency and the phase of the signal, the time-frequency analysis is used. In this way, pure signal and noise can be achieved separately since the noise is usually condensed in a specific region of the frequency domain [83]. In DL, features are extracted automatically. Different time-frequency representations affect deep learning performances as they can highlight different hidden features. Therefore, in our study, ECG signals were transferred to 2D using different time-frequency representations and examined with different CNN models.

3.3 Classifiers

Classification techniques vary for supervised, semi-supervised and unsupervised machine learning models. In this study, eight classifiers, which are explained in the following subsections, are used in supervised learning models.

In supervised learning models, each data has a label. Labels and data are trained together. These labels are activity names and subject IDs for our studies. After the model is trained, it is expected to estimate the labels for the new input data. During biometric verification, the main process is whether a person is the person she/he claims to be. That is enrollment samples with known a label is compared with test samples. For this reason, supervised learning methods are used in our experiments.

In classical Machine Learning (ML), classifiers and their optimizers have a great impact on the classification results. Situations, where each classifier is advantageous or disadvantageous, may occur depending on the features used. The pros and cons of each classifier are explained in subsections. To summarize, Linear Discriminant Analysis (LDA) are linear classifiers. The LDA operates on the assumption of Gaussian circumstantial density models [210]. The LDA requires the data to be normally distributed and it can also eliminate outliers. The k-Nearest Neighbours (KNN) is a non-parametric classifier that has a higher computational

complexity than the NB for large datasets. The LDA, on the other hand, is a fast, simple and easy-to-use classifier [211]. The NB is an optimal choice to minimize computation costs [212]. The NB utilises the prior class probabilities for the test samples and performs well unless it faces the issue of zero probability. The Decision Trees (DT) is a non-linear classifier that can effectively address classification and regression problems. It also contributes to reducing computation costs and exhibits more robustness to missing values and non-normalized data [1].

These classifiers were chosen because they are frequently used in ML models for ECG-based applications [213]. In our studies, eight classifiers were used because the effect of the generated new features and the physical activity/emotion status on biometric verification were investigated.

3.3.1 K-Nearest Neighbour (KNN)

A KNN classifier in a non-parametric model group allocates a class to each test sample based on the majority class of its k neighbours [214, 215]. The only hyper-parameter is the number of k and each test sample class assigns using the closest distance measurements between the k neighbours training sample and testing sample. Choosing the k value depends entirely on the structure of the dataset. In order to avoid over-fitting or under-fitting problems, the value k should be selected correctly. The k value is usually an odd number and if the input signal is noisy, a high k value should be selected. In our experiments, in order to see the effects of different k values on the extracted fiducial features (cf. subsection 3.2.2.1), k=*3*, k=*5*, k=*7* and k=*9* values were examined in the biometric verification model. In addition, the KNN classifier was used for activity classification with Manhattan distance and Euclidean distance-based features (cf. subsection 3.2.2.2) when k=*3*.

Euclidean (Eq.(3.1)), Manhattan (Eq.(3.2)), Minkowski (Eq.(3.3)) and Hamming (Eq.(3.4)) distance functions are commonly used for distance measurements.

KNN is known as a weak classifier as it performs insufficiently in cases of increasing data and features. However, it is a simple and easily adaptable algorithm and it gives acceptable accuracy for a small dataset [180].

3.3.2 Subspace KNN Ensemble (S-KNN)

A random subspaces ensemble method with KNN and DT was introduced by Ho [216] to reduce the high dimensionality and high computational cost problems in the KNN model.

The random subspace method is a stochastic process. In this process, samples in the given feature vector are randomly selected to build each classifier. The selected samples in the different subspaces are compared with test samples. The distances between test samples and samples on the each subspace are calculated by different KNNs. A majority vote of each KNN is calculated and a final decision is made. The outline of the S-KNN classification is shown in Fig.3.7. In this figure, *F1, F2* and *Fx* represent features in training data. *S1, S2* and S_n show randomly selected subspaces. *D1, D2* and D_n represent the decision of each KNN classifier.

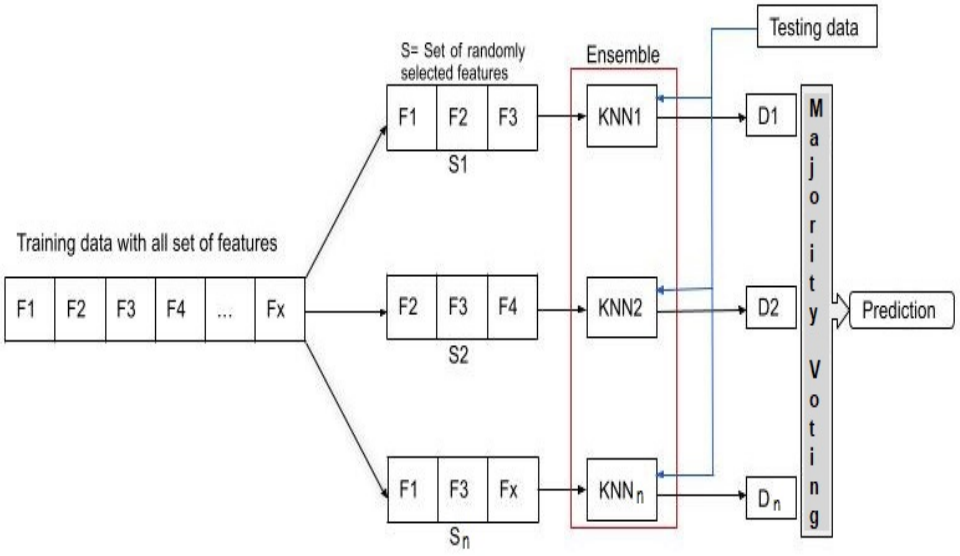

FIGURE 3.7: The outline of the S-KNN classification

A *Classification Learner* toolbox was used for creating a classification *Model* in MatLab R2020a. The S-KNN classifier was used in the activity classification

case to compare the performances of new features and the effects of different training/testing rates.

3.3.3 Naive Bayes (NB)

A Gaussian NB has been employed for two-class and multi-class classification problems. The input of the NB system can be binary or categorical. The system is based on the Gaussian distribution of the training data and class probability calculations.

Instead of trying to calculate the values of each feature, given the target value they are assumed to be conditionally independent. The system is based on only prior class probabilities for test samples, for this reason, computational costs are generally lower than other classifiers. The system assumes a probability for each class and each feature's conditional probability for belonging to each class (i.e. *a-prior* probabilities) at the beginning of the training process. Choosing the *a-posterior maximum probability* is the only decision rule in the NB classifier [66].

When C is a class and F is a feature, $P(C\,|F)$ is the class posterior probability according to the predictor, $P(F)$ and $P(C)$ are the prior probability of the predictor and the class respectively, the NB theorem can be explained as Eq.(3.7).

$$P(C|F) = P(F|C) * P(C)/P(F) \tag{3.7}$$

The NB classifier is useful for solving multi-class and two-class classification problems with small amounts of training data, and categorical data. However, it can encounter a zero-probability problem, especially for categorical inputs. In addition, the system assumes all features are independent which may be problematic in a real life scenario. In our studies, the NB classifier was used in the biometric verification case with different sets of features and several genuine/imposter sample rates.

3.3.4 Linear Discriminant Analysis (LDA)

A linear combination of features is found by LDA to characterize or distinguish classes using sample distributions. LDA can be used for multi-class or two-class classification problems.

LDA assumes that each class has a normal distribution and all classes have a similar covariance matrix. If the data does not conform to these assumptions, the system cannot function properly. LDA works with the covariance of the features unlike, NB which works with the assumptions of conditionally independent features. In LDA, decision boundaries can be found by observing features where the large mean differences and small variances within classes occur. For this reason, the training set size, data normalisation processes and outliers in the data might affect system performance.

To understand the LDA classification: Given a random variable X that belongs to one of C classes, each with a class-specific probability density function *f(x)*, a discriminant rule aims to partition the data space into C mutually exclusive regions corresponding to the classes. With these regions, discriminant analysis performs classification by assigning x to the class k if x lies in the region k. Assuming that X follows a multivariate Gaussian distribution with mean vector (μ) and covariance matrix (Σ), the LDA criterion can be derived as Eq.(3.8) while π represents the class prior probabilities. Based on the Bayesian rule, we assign the data x to class k if it has the maximum posterior probability among all C classes for $i = 1,\ldots,$ C.

$$\delta_i(x) = log(f_i(x)) + log(\pi_i) \tag{3.8}$$

In our studies, the LDA is used for biometric verification purposes with fiducial features to see the effects of different genuine/imposter rates on the wearable and medical ECG recorders.

3.3.5 Support Vector Machines (SVM)

SVM algorithm can be implemented as either a parametric or non-parametric model, contingent upon the kernel function employed. It transfers the classification problem to an optimization problem to reduce computational costs [217].

SVM can be explained with three main elements: support vectors, hyperplanes and margins. A hyperplane is a decision plane used to divide classes. The closest training data points to the hyperplane are called support vectors. A margin is calculated from the perpendicular distance between support vectors. In an ideal classification scenario, the margin would be large. An example of an SVM implementation is shown in Fig. 3.8.

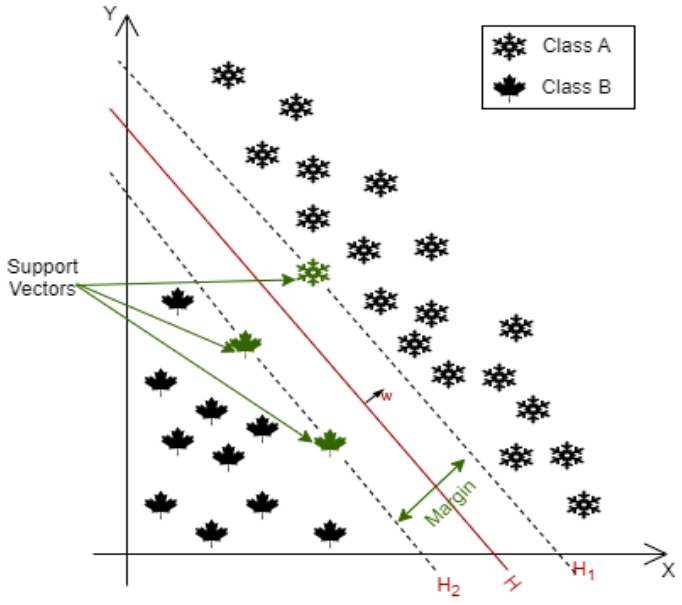

FIGURE 3.8: An example SVM implementation

In Fig. 3.8, H represents an optimal hyperplane or classification line. When the Class A side is assumed as a positive (+1) class, the Class B side will be a negative (-1) class. In that case, H_1 and H_2 support vectors and H can be represented as Eq.(3.9), Eq.(3.10) and Eq.(3.11), respectively.

$$H_1 : \{x|(w + x)b = +1\} \tag{3.9}$$

$$H_2 : \{x|(w + x)b = -1\} \tag{3.10}$$

$$H : \{x|(w + x)b = 0\} \tag{3.11}$$

In these equations, x represents the corresponding label of input data, w is the normal vector of the hyperplane while b is a constant value [214].

The SVM classifier is a useful option for high dimensional data with a large amount of features. Many kernel structures such as linear, non-linear and sigmoid can be used within the SVM. However, finding optimal parameters might be problematic because it is difficult to find a good kernel function that can map data into a higher-dimensional space where it can be linearly separated. The choice of kernel function can have a significant impact on the performance of SVMs. In addition, computational costs might be high with non-linear kernel structures in the training. It is generally useful for small datasets because of the non-scalable nature of the SVM.

Over-fitting is a common problem in ML algorithms. In SVMs, over-fitting can occur when the model is too complex or when the size of the training data is too small. Over-fitting occurs when the model is trained on a small amount of data and it becomes too specific to that data. To eliminate the over-fitting, adding structural constraints on the discriminant surface can be added.

3.3.6 Decision Tree (DT)

DTs are also a non-parametric model. The depth of the tree (the number of branches and nodes) is a hyper-parameter of this classifier. It learns all probabilities and conditions by dividing the given feature vector into parts [218]. This is usually a grouping of samples with the same label. Learning starts at the root of the tree, continues through the branches, and the classification is contained at a leaf node. Each branch has its own threshold which is conjectured by the previous

branch's input features to divide groups. DTs are implemented as *if statements* to make a decision. If an input value is larger than a threshold, it is attributed to group A, if an input value is smaller than a threshold, it is attributed to group B.

Using DTs enables an understanding of the Boolean logic through the visual representation of the trees, unlike a Neural Network. Furthermore, it is useful for many data types such as continuous and discrete data. For this reason, it can overcome missing value problems, unlike Naive Bayes. DTs are not sensitive to correlated features (i.e. if inputs have two similar features, the algorithm will select one of the features to split.) to reduce computational costs [219]. On the other hand, DTs can be vulnerable to over-fitting, variations in the data and high computational costs in training. DTs can overfit the training data and lose their generalization capabilities. In this case, while the training data is well classified, errors increase in the classification of test data. The over-fitting problem can be solved by pre-pruning (i.e. reducing the number of nodes in DTs) and post-pruning (i.e. removing nodes) processes.

3.3.7 Bagged Tree Ensemble (BT)

The bagging method which, was introduced by Leo Breiman [220], was used across seven different datasets and it was observed that the rate of misclassification was reduced with this method. Bagging is an ensemble learning technique that enhances the performance and robustness of machine learning models by utilising multiple models instead of a single one. It operates by creating diverse samples of data from the original training set by sampling with replacement. Then, it trains a distinct model on each sample and aggregates their predictions by averaging (for regression) or voting (for classification). Bagging can reduce the variance of models susceptible to over-fitting, such as DTs, and improve their generalization ability. The general implementation of BT is shown in Fig.3.9.

A single DT model can exhibit low bias and high variance issues, implying that it can fit the training data very closely but fail to generalize to new data. This

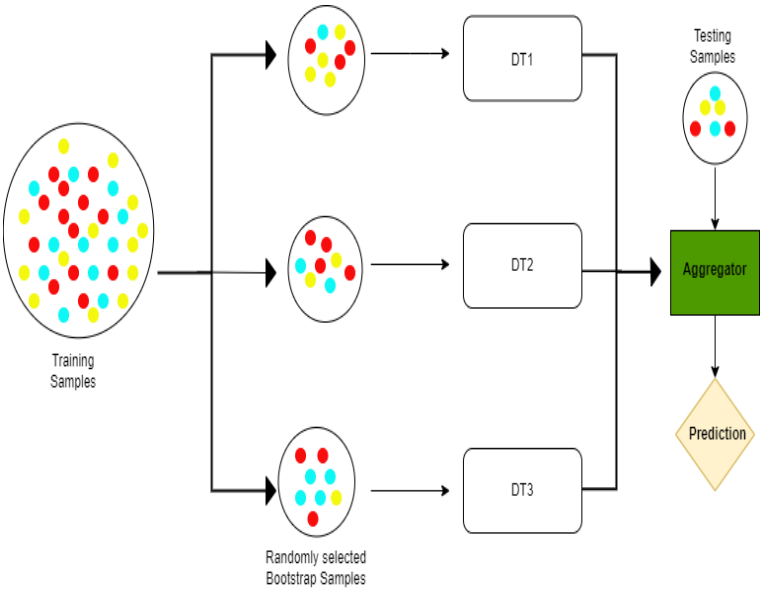

FIGURE 3.9: An example BT implementation

is because a single DT can grow very deep and complex, capturing the noise and outliers in the data. Bagging can help to mitigate this issue by averaging the predictions of multiple DTs that are trained on diverse samples of data, resulting in a more stable and accurate model [221]. Reducing the variance can be helpful, especially in high-dimensional datasets. However, if the number of iterations grows in the classification processes, the computational costs will increase.

3.3.8 Neural Networks (NN)

Artificial Neural Networks or NN tries to copy the human brain (i.e. brain neurons) within a computer environment. It contains an input layer, many hidden layers and an output layer. Each layer has many artificial neurons which are connected to other neurons. These connections have their own weights and threshold to activate the neurons and data transfer.

In a NN, each neuron computes a weighted sum of its inputs and applies an activation function to produce an output. The weights are parameters that determine how much each input contributes to the output, and the activation function introduces non-linearity to the network. The weights are not fixed but are learned

from the training data using an optimization algorithm such as gradient descent. The optimization algorithm adjusts the weights based on the error between the predicted output and the actual output, with the goal of minimizing the error. The thresholds are also parameters that determine when a neuron is activated or not, and they are also learned from the data. However, the thresholds are usually represented as bias terms that are added to the weighted sum of inputs, rather than subtracted from it.

NNs have the ability to learn and model complex relationships between inputs and outputs. For data classification, some data sets might be linearly separable, meaning that a straight line can divide the classes, but some might be nonlinearly separable, meaning that a more complex boundary is needed. Neural networks can handle both linear and nonlinear data sets by using adjustable weights and activation functions that introduce non-linearity to the network. However, neural networks are not the only models that can manage complex data. Other models, such as the SVM or KNN, can also deal with nonlinear data sets by using different kernels or distance metrics. However, we used the NN and SVM classifiers as linear models in our studies.

Chapter 4

Phase 1: ECG Biometric Verification Across Activities: Direct Biometric Verification

As stated in Chapter 2, this study focuses on the comparison of multiple factors in ECG biometrics using ML and DL models. This chapter concerns the key issues of ECG data that contain physical activities and emotional states for biometric verification. The first issue addressed concerns the quantity of enrollment data. This is addressed in subsection 4.1.1. The second issue to be explored concerns comparing the biometric verification performances of fiducial features, time interval features, Manhattan-based features and Euclidean-based features in terms of EERs on medical and wearable devices and databases. This is addressed in subsection 4.1.2. A final experiment is Section 4.2 presents a comparative analysis of the impact of various ECG representations on biometric systems, taking into account differing deep learning hyperparameters, enrollment durations, and devices. These topics will be assessed by employing ML and DL models in Section 4.1 and Section 4.2, respectively.

4.1 Machine Learning Models

Across ML models, we have two different approaches. The first approach is to vary enrollment time with common ECG features with a second approach creating new features and assessing performances with different classifiers.

4.1.1 Enrollment time effects on biometric verification

In this section, E-HOL and WeSAD datasets are used to investigate the effects of enrollment time and different classifiers on biometric verification.

Enrollment (i.e. training) sample sizes are generally proportional to verification rates. A higher number of training samples per subject generally yields higher verification rates. The objective of this study is to propose a robust verification system with various training sample sizes (i.e. 5, 15, 25 and 50 genuine samples) for medical and wearable-based ECG data. Previous studies have shown promising results. The diversity of data is crucial for the validity and wide applicability of authentication systems. However, previous studies generally lack a realistic scenario, and in most cases, a small number of subjects participated in the study, resulting in over-fitting and unreliable data. The proposed method achieves consistent verification results for both the medical and wearable ECG data, with few training samples and a short enrollment time, which are essential factors for a commercial biometric authentication system.

The proposed ML models used 10-second windows for each sample and multiple features described in subsection 3.2.2.1 calculated from each sample. LDA, DT, NB and KNN classifiers were utilised for binary classification in biometric verification tasks. A random subject was selected as a genuine and imposter samples were randomly selected from remaining subjects. 50, 150, 250 and 500 seconds (i.e. 5, 15, 25 and 50 genuine samples per subjects) from the sample chosen among each genuine and imposter subjects were utilised during the training phase to construct model. The genuine and imposter samples were selected randomly and in equal

numbers. Thus, they could represent any possible activity or emotional state. To increase the reliability of the system and to avoid issues such as over-fitting (this may happen if the samples were derived from the same activity or emotional state) or under-fitting (this could happen if the samples were obtained from different activity or emotional states), at least 5 genuine samples (i.e. *5 genuine samples * 10-second windows = 50 seconds enrollment time*) were chosen for the training phase. Training imposter samples were randomly drawn from the subjects in the dataset who were not enrolled. In addition, to increase the reliability of the model, instead of testing only 1 genuine subject, different subjects were randomly selected as a genuine (200 subjects for EHOL and 15 for WeSAD) and the mean EER was calculated. The biometric verification time (i.e. testing phase) was equivalent to the duration of a single sample window (i.e. 10 seconds).

The outcomes are presented as EER values for each enrollment set, model and dataset. This study employed classifiers for binary classification. Specifically, the genuine samples were assigned the class *"1"* and the imposter samples the class *"0"*. The classifier performed separate classification for 200 subjects (for the E-HOL dataset) and 15 subjects (for the WeSAD dataset). The mean EER values are presented in Table 4.1 for the E-HOL dataset and Table 4.2 for the WeSAD dataset using only fiducial and time interval features.

TABLE 4.1: The biometric verification results in terms of Equal Error Rates(%) for the E-HOL dataset.

E-HOL Dataset				
Classifiers	**Enrollment Time (seconds)**			
	50s	**150s**	**250s**	**500s**
KNN(k=3)	39.32%	38.69%	37.80%	32.01%
KNN(k=5)	43.53%	43.25%	40.25%	35.47%
KNN(k=7)	44.85%	44.72%	42.59%	37.34%
KNN(k=9)	45.67%	42.96%	38.37%	37.89%
LDA	37.33%	25.65%	19.22%	15.94%
DT	27.91%	16.64%	10.99%	7.89%
NB	6.30%	4.17%	4.09%	3.64%

As shown in Table 4.1, the EER values decrease with increasing enrollment time. The KNN classifier yielded the highest EER values, while the NB classifier was the most effective. The EER values for all conditions increased with higher KNN

k values. The KNN classifier requires a larger k value to achieve more accurate decisions in noisy data. These results indicate that our signal filtering method was effective as we obtained better results with a low k value.

TABLE 4.2: The biometric verification results in terms of Equal Error Rates(%) for the WeSAD dataset.

WeSAD Dataset				
Classifiers	**Enrollment Time (seconds)**			
	50s	**150s**	**250s**	**500s**
KNN(k=3)	45.98%	37.63%	35.04%	31.94%
KNN(k=5)	46.7%	37.53%	35.41%	32.63%
KNN(k=7)	48.79%	38.55%	38.02%	32.05%
KNN(k=9)	47.36%	40.57%	37.08%	35.62%
LDA	32.03%	16.10%	11.69%	9.97%
DT	31.88%	9.61%	8.37%	1.60%
NB	4.57%	3.67%	3.67%	3.02%

According to the results in Table 4.2, the EER distribution is similar to that in Table 4.1, and the EER decreases with increasing enrollment time. A smaller k value generally yielded lower EER results. The NB classifier outperformed the others, while the DT classifier achieved the lowest EER value of 1.60% for the 500 seconds training condition.

The EER values for the WeSAD dataset are generally lower than those for the E-HOL dataset. This can be attributed to the smaller number of subjects and shorter recording time in the WeSAD dataset. Moreover, although a direct comparison is not feasible, while physical activities are limited and different emotional states are measured during ECG recording in the WeSAD dataset, neither emotional states nor physical activities are limited in the E-HOL dataset. For this reason, the changes in the E-HOL dataset were greater, which may have affected the results.

One possible explanation for the lower EERs of NB and DT over KNN in a biometric verification task is that KNN suffers from the drawbacks of being sensitive to noise and outliers in the data set, which can impair its predictive ability. In contrast, NB and DT may exhibit more resilience to these challenges. Another possible explanation is that the KNN belongs to the class of lazy learners, which defer the learning process until new data is presented for prediction [222]. On the

other hand, NB and DT are examples of eager learners, which construct a model from the training data prior to prediction [223]. This can confer an advantage to NB and DT in terms of speed and accuracy on large or complex data sets.

4.1.2 Feature effects on biometric verification

In this section, Vollmer and WeSAD datasets containing ECG data collected during physical activities and different emotional states were used. Features based on Manhattan distances and Euclidean distances were extracted from ECG data consisting of 4 second time windows and their effects on biometric verification with different classifiers were investigated. In Phase 1, overall biometric verification performances were evaluated irrespective of the physical activities and/or emotional states conducted. The outline of the biometric verification process is shown in Fig.4.1.

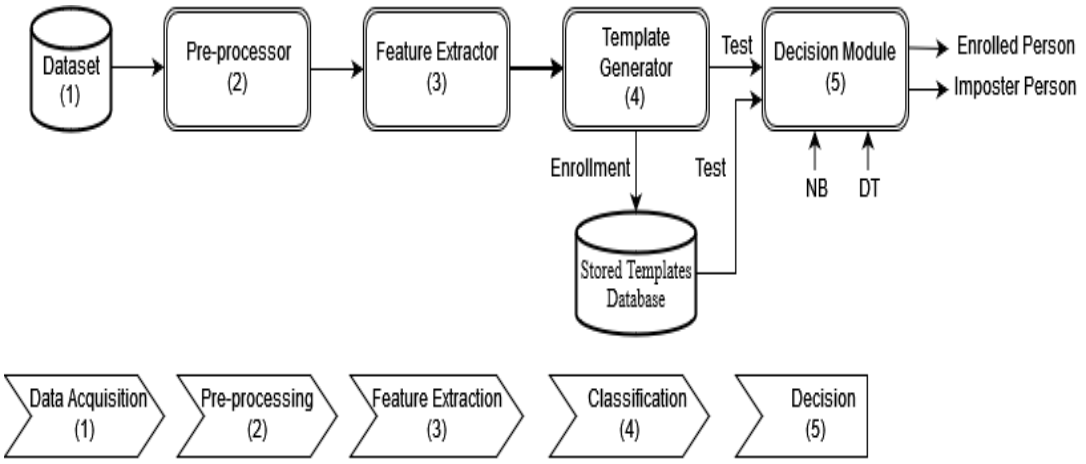

FIGURE 4.1: Biometric verification model is shown at the top and the classification steps for biometric verification are shown below.

The data from each device underwent separate testing. Data were divided into two subsets: training and testing. The training subset comprised the data of 12 out of 15 subjects from the WeSAD dataset and 10 out of 13 subjects from Vollmer's dataset. The testing subset consisted of the data of 3 unseen subjects selected from either dataset. In each trial, the unseen subjects were varied and

each subject appeared in the testing subset at least once (i.e. the subjects selected for testing were altered sequentially, such as subjects 1-2-3, 4-5-6, and so on).

The number of genuine samples was determined by the number of samples per person. Not all samples collected from the training subjects were utilised for training purposes to reduce computational costs and training time. Similarly, not all samples obtained from subjects assigned as unseen subjects were utilised for testing. For example, to investigate the impact of the quantity of training and testing samples on biometric verification, we used 70% of the subject samples for training and used 30% of the samples of the unseen subject for testing. This example defines M-30% for features based on Manhattan distance and E-30% for features based on Euclidean distance. Imposter samples were randomly chosen from the remaining unused samples. To compare results, both Manhattan and Euclidean distance features also used 40% and 50% of the genuine sample size in the testing set (i.e. M-40%, M-50%, E-40%, E-50%). The number of genuine and imposter samples was equal for both training and testing data. The same ratio of genuine samples for all datasets and devices was used to achieve comparable results. The number of genuine samples per activity was approximately equal. However, each dataset differed in the number of subjects, samples per device, activities and data collection procedures. The number of each subject's genuine samples for training and testing is shown in Table 4.3.

TABLE 4.3: The number of genuine samples in testing and training per person.

Genuine Samples		**RespiBAN**	**Faros**	**SomnoTouch**	**Nexus-10**	**Hexoskin**
50%	Train	534	731	770	737	837
	Test	534	731	770	737	837
40%	Train	641	878	924	876	1005
	Test	427	584	616	589	669
30%	Train	748	1024	1078	1032	1172
	Test	320	438	462	442	502

The first step was to use the biometric verification method to evaluate the performance of the new feature set. In this step, no activity classification system was

involved. Therefore, genuine and imposter samples were randomly chosen according to their ratios of 30%, 40% and 50%. Fig.4.2 and Fig.4.3 show the NB and DT classifier performances for the RespiBan device from the WeSAD dataset and the Faros, SomnoTouch, Nexus-10 and Hexoskin devices from the Vollmer dataset, respectively.

FIGURE 4.2: EERs(%) were shown according to the NB classifier with 50%, 40% and 30% genuine samples in test data. *M* is Manhattan distance-based features and *E* is Euclidean distance-based features.

Fig.4.2 illustrates the results of Manhattan distance-based features with 30% of samples from unseen participants in testing having the lowest EERs (i.e. M-30%). The distribution of FAR and FRR with this testing set size is indicated. In M-30% cases, the RespiBAN device has a FAR of 20.03% and FRR of 19.59%, the Faros device has a FAR of 17.67% and FRR of 17.05%, the Somnotouch has a FAR of 20.87% and FRR of 19.21%, the Nexus has FAR of 18.53% and FRR of 18.71%, and the Hexoskin device has FAR of 17.69% and FRR of 18.85%. The Euclidean distance-based features show the same trends as the corresponding Manhattan feature sets. As the training set sizes increase, EER decrease and the lowest EER are obtained from E-30% cases.

According to results in Fig.4.3, from M-50% to M-30% and from E-50% to E-30%, EERs decrease for all devices, as can be observed. A large number of testing samples leads to higher EERs.

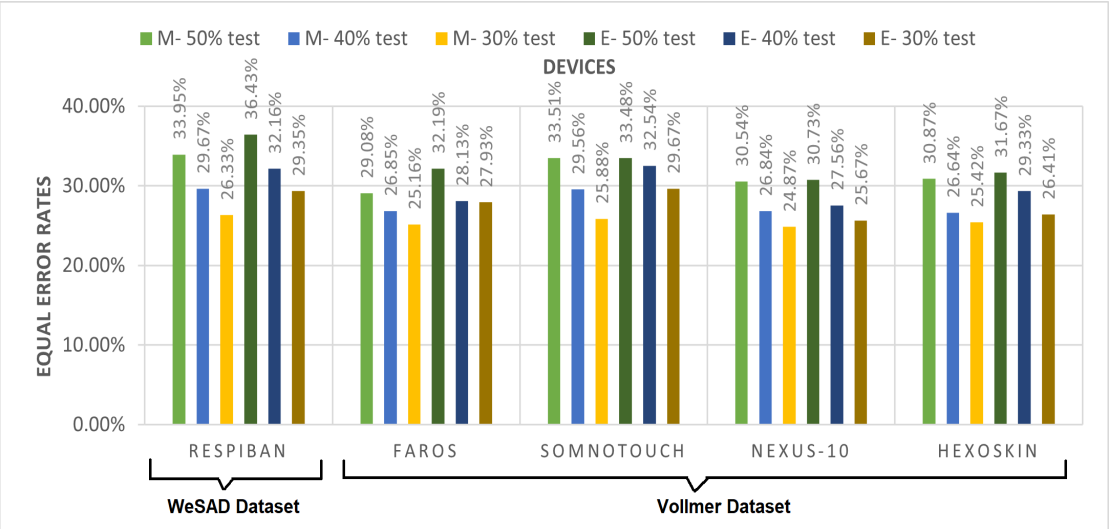

FIGURE 4.3: EERs(%) were shown according to the DT classifier with 50%, 40% and 30% genuine samples in test data. *M* is Manhattan distance-based features and *E* is Euclidean distance-based features.

A general trend observed in Fig.4.2 and Fig.4.3 is that the feature sets that employed Manhattan distances surpassed those that utilised Euclidean distances. The NB classifier exhibited lower EERs in all instances, but the trends were consistent for both NB and DT. The data derived from RespiBAN and SomnoTouch devices yielded higher EERs. The Faros, Nexus-10 and Hexoskin data attained similar performances. Despite being a consumer-based device, the Hexoskin demonstrated verification performances analogous to other medically-approved devices. The lowest EERs for each testing size (i.e., 50%, 40%, and 30%) were obtained from different devices. For instance, while the lowest EER of the M-40% case was achieved from the Hexoskin device in Fig.4.3, the Nexus-10 device had better EERs in 30% and 40% sizes than other devices. Furthermore, while the lowest EER of the E-30% case was obtained from the Faros device in Fig.4.2, the Nexus-10 device had lower EERs in E-50% and E-40%. This implies that device selection, feature selection, and training/testing sizes affect biometric verification performance.

4.2 Deep Learning Models

As described in Section 4.1, a comprehensive analysis of various factors influencing biometric verification performance in classical ML models was conducted. Subsequent investigations were conducted to ascertain the impact of varying image representations and enrollment times on DL models applied to medical and wearable devices for biometric verification purposes. The image representations generated for the DL model and the selected time intervals were described in Section 3.2.3. In this section, we used ResNet50 and DenseNet201 CNN models on the data from the Faros and Hexoskin devices from the Vollmer dataset.

In the context of DL, it is imperative to utilise both validation and testing sets independently in order to mitigate over-fitting and accurately assess the model [224]. The validation set serves to evaluate the model during training and facilitate hyperparameter tuning [225]. Conversely, the test set is employed to measure the final performance of the model on previously unseen data and facilitate comparisons with alternative models. This partitioning enhances the model's capacity to generalize on unseen data [225]. Table 4.4 describes the number of images in 2, 4 and 10 seconds time windows utilised for training, validation and testing for both CNN structures.

TABLE 4.4: The number of images in training, validation and testing sets

	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

ResNet and DenseNet are both widely utilised DL architectures that exhibit competitive performance. Nevertheless, both architectures possess inherent limitations. In the case of ResNet, the shortcut connection among convolutional blocks employed to stabilize training may also constrain its representational capacity.

Conversely, DenseNet possesses a higher capacity due to its utilisation of multi-layer feature concatenation. However, this dense concatenation introduces the issue of increased GPU memory and training time requirements [226]. 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.

In the process of training a CNN using MATLAB, 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 [227].

The mini-batch size dictates the number of training samples utilised in each iteration of the optimization algorithm. While a smaller batch size may converge more rapidly than a larger batch size, a larger batch size may attain optimal minima that are unattainable with a smaller batch size. Additionally, a smaller batch size may have a substantial regularization effect due to its high variance, necessitating a smaller learning rate to prevent overshooting the minima [228].

The initial learning rate determines the magnitude of the step taken by the optimization algorithm at each iteration. A larger learning rate may result in more rapid convergence, but it may also cause the optimization algorithm to overshoot the minima and fail to converge. Conversely, a smaller learning rate may result in more stable convergence but may necessitate additional iterations to reach the minima.

Considering the aforementioned effects and criticalities, we explored the optimal hyperparameters manually. As a starting point, we used a set of default parameters in MatLab and changed it according to our tasks using the ResNet50 model. For instance, an initial rate of 0.01 and a mini-batch size of 128 were used for 3 trials using *Sc4, Mel4* and *Sp4* cases which had the lowest EER results overall. However, the classification accuracy rates for validation were insufficient. Then, we reduced the mini-batch size to 64 for the same 3 image types. As the same in 128 mini-batch size, when 64 mini-batch size was used for the same cases, the validation accuracy rates and the validation loss outputs obtained for each epoch could not be obtained. To increase the validation accuracy rates, we decreased initial learning

rates to 0.0001 as the study [83] and a mini-batch size of 16. It has been found that the execution time for training is longer when the parameters are used. 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. In all cases, training and validation accuracy rates were higher than 90%. For this reason, the number of epochs has not been increased to avoid over-fitting.

In the following subsections, ResNet50 and DenseNet201 CNN models created for biometric verification and the results obtained will be explained.

4.2.1 ResNet50 CNN

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. [229], 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. The ResNet50 CNN was used to reduce time spent on training and biometric verification and to achieve a fast result without significantly compromising performance.

The original ResNet architecture consists of 34 layers with shortcut connections that skip some layers, transforming a regular network into a residual network. This alleviates the vanishing gradient problem that occurs when the network becomes deeper. The vanishing gradient problem arises during the training of an ANN and CNN when each weight in the network receives an update proportional to the partial derivative of the error function with respect to the current weight [230]. In some instances, the gradient may become infinitely small, effectively precluding any alteration in the weight's value and potentially impeding further training of the neural network [230]. Each convolutional layer has 3x3 filters. ResNet50, however, adopts the bottleneck block structure, which reduces the number of parameters and

matrix multiplications using 1x1 convolutions [231]. This structure also enables faster training. It has three layers instead of two in each block. Although the ResNet architecture addresses the vanishing gradient problem, it may incur more computational costs due to the additional layers. Fig.4.4 illustrates the details of the ResNet50 CNN model used for training and biometric verification. The output size at each stage is indicated as "$n \times n$" after each block. In addition, each identity block is an example of the bottleneck block structure.

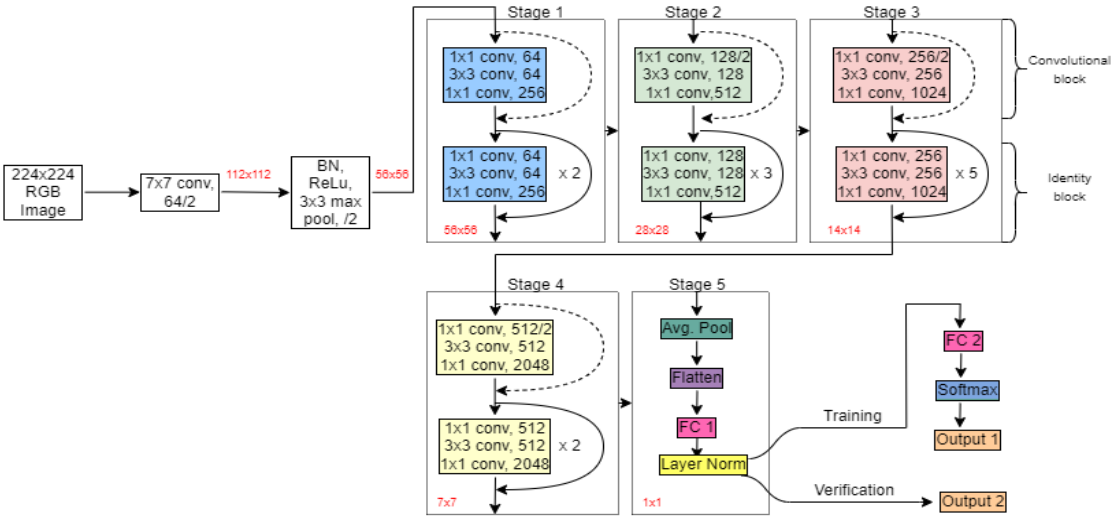

FIGURE 4.4: Used ResNet50 model in biometric verification

The output of each stage is used as the input of the next stage. A *Convolutional block* represented in Stages 1, 2, 3 and 4 have 3 convolutional layers. Every convolutional layer is followed the Batch Normalization (BN) layer and the ReLu layer. In addition, a parallel branch, which is represented by a dashed line in Fig.4.4 has a convolutional layer and a BN layer. An *Identitiy block* shares the same architecture as the convolutional block, except that the shortcut connection (i.e. a parallel branch) does not contain any convolutional or BN layers.

In the training phase, we used the ResNet50 architecture and trained it from scratch. The data from 11 subjects out of 13 were randomly split into 80% training set and 20% validation set (i.e. it is a fixed split, no cross validation), while the data from the remaining 2 subjects were reserved for biometric verification (i.e. unseen subjects). Unseen subject pairs were selected as *P1-P2, P6-P7, P8-P9* and *P11-P12*. 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.

During the training phase, the data of 11 subjects were labelled as 11 different classes, and biometric identification was performed in the training and validation phases. During the training phase, each time-frequency representation was examined in different cases. For example, in the spectrogram case, only spectrogram images were given to the model for training and the results were examined. For the scalogram case, only scalogram images were fed to the CNN model, and for the mel-spectrogram case, only Mel-spectrogram images were fed to the CNN model.

The system is retrained for each pair of subjects, as the samples from the remaining 11 subjects will vary for each pair of unseen subjects. For instance, where *P1-P2* is selected as the unseen subject pair, the other 11 subjects (from P3 to P13) are used for training and validation. When it is necessary to select another set of unseen subjects (e.g. *P6-P7* subjects) after all required results have been obtained, the model is trained with the remaining subjects (i.e. from P1 to P5 and from P8 to P13). This ensures that the model fits new unseen subject pairs. The *Flatten, FC1* (Fully Connected) and *layer normalization* layers are used to produce embeddings. 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 this way, *Output1* gives us biometric identification results for 11 subjects (in terms of percentage), while *Output2* gives us embedding vectors required for biometric verification. The parameters used for training and the number of training and validation images used are described in Section 4.2.

FC layers constitute a fundamental component of CNNs [232]. The output generated by the convolutional layers represents high-level features inherent in the data. Although this output could be flattened and directly connected to the output layer, the incorporation of the FC layer generally provides an inexpensive

method for learning non-linear combinations of these features. In essence, the convolutional layers furnish a meaningful, low-dimensional, and relatively invariant feature space, while the FC layer learns a function, which may be non-linear, within this space. Layer normalization constitutes a technique employed in the field of deep learning to standardize the inputs to a specific layer for each individual training sample. This process can enhance the stability of the neural network, expedite the training process, and mitigate the occurrence of over-fitting [233]. In classification tasks, the *Softmax* layer is commonly employed as the terminal layer of a CNN architecture. The *Softmax* function is applied to the output of the preceding fully connected layer, yielding a probability distribution over the predicted output classes. The class with the highest probability is designated as the predicted class, and the output of the *Softmax* layer represents the likelihood of a given input belonging to a specific class [234]. These layers are incorporated into the original ResNet architecture. This is due to the fact that the layers employed to train the system, as well as the biometric verification task executed using the trained system, yield disparate outputs with distinct FC layers.

Upon completion of the training process, a *DAGNetwork* (Directed Acyclic Graph Network) is constructed. this is a neural network architecture utilised in DL wherein the layers are structured as a directed acyclic graph. This configuration facilitates the construction of more intricate architectures, wherein individual layers may receive inputs from and transmit outputs to multiple layers. DAG-Networks are versatile and can be employed in a diverse array of tasks, including image classification, object detection, and semantic segmentation [235, 236]. This *DAGNetwork* is subsequently saved and the final three layers are removed. As depicted in Fig.4.4, 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. Essentially, the CNN model is trained to generate embeddings for each image. Then, a predetermined number of images

(N= 1, 3 or 5) for each subject are selected for enrollment. The average of embeddings which were obtained from N images for an enrollment is considered the template for the subject. This relationship can be mathematically represented as depicted in Eq.(4.1). In this formula, $(X_i)_s$ represents i_{th} enrollment image of the selected subject s. M symbolizes the created embedding model while T_s is the template for the selected subject [66].

$$T_s = \frac{1}{N} \sum_{i=1}^{N} M\left((X_i)_s\right) \tag{4.1}$$

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 minimizes the combined error.

Mean EERs are shown for different numbers of genuine samples and different image representations (i.e. scalogram, spectrogram, Mel-spectrogram) in Table 4.5. *Sc2, Sc4 and Sc10* are scalogram images that were created from 2 seconds, 4 seconds and 10 seconds time windows. *Sp2, Sp4, and Sp10* are spectrogram images and *Mel2, Mel4, and Mel10* represent Mel-spectrogram images with the same time windows. *1-S, 3-S,* and *5-S* represent 1, 3 and 5 genuine samples used in enrollment. This means that the enrollment time for time windows of 2 seconds is 2 (in 1-S), 6 (in 3-S), and 10 (in 5-S) seconds respectively. Similarly, for time windows of 4 seconds, the enrollment time is 4, 12, and 20 seconds respectively. For time windows of 10 seconds, the enrollment time is 10, 30, and 50 seconds, respectively.

A general trend was observed in which the EER decreased with an increase in the number of genuine samples. Furthermore, the biometric verification performance

TABLE 4.5: Mean EERs(%) are shown for different numbers of genuine samples and three image representations. *Sc2, Sc4 and Sc10*: Scalograms. *Sp2, Sp4 and Sp10*: Spectrograms. *Mel2, Mel4 and Mel10*: Mel-spectrograms in 2 seconds, 4 seconds and 10 seconds time windows. *1-S, 3-S, and 5-S*: The number of genuine samples in enrollment.

Devices	Faros			Hexoskin		
# of Genuine Samples / Image Types	1-S	3-S	5-S	1-S	3-S	5-S
Sc2	37.57%	33.28%	32.95%	41.48%	36.45%	34.46%
Sc4	**15.93%**	**13.42%**	**11.15%**	34.01%	**12.19%**	**11.21%**
Sc10	22.39%	19.78%	18.96%	**22.53%**	20.47%	19.09%
Sp2	38.09%	38.53%	37.35%	39.11%	33.28%	34.64%
Sp4	34.97%	30.64%	28.62%	35.97%	30.64%	28.62%
Sp10	27.06%	26.83%	26.96%	24.39%	20.95%	21.23%
Mel2	38.71%	37.68%	38.67%	38.34%	39.22%	39.48%
Mel4	39.64%	37.99%	37.75%	35.50%	34.27%	35.68%
Mel10	32.83%	32.69%	30.08%	28.10%	28.79%	27.55%

of the Faros medical device surpassed that of the Hexoskin consumer-based device. Scalogram images generally yielded the lowest EERs, while Mel-spectrograms obtained the highest EERs. An increase in the time window length (from 2 seconds to 10 seconds) for spectrogram and Mel-spectrogram images resulted in lower EERs, whereas the minimum EERs for scalogram images were obtained with a 4 seconds 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 extracted information from the images and the number of images which was used in the training should be balanced. Heartbeats are clearly visible in scalogram images (see Fig.3.6). Although images consisting of 10 seconds time windows contain more heartbeats than 4-second images, the total number of images obtained per subject is less (see Table 4.4). Although this provides more detail in the image, it causes the model to be trained with the fewer number of images.

4.2.2 DenseNet201 CNN

DenseNet is utilised to address the vanishing gradient problem, similar to ResNet. 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. This means that the results of each layer serve as inputs for subsequent layers [237]. 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.

Fig.4.5 illustrates the details of the DenseNet201 CNN model used for training and biometric verification. The output size at each stage is indicated as "$n \times n$" after each block. In addition, certain structures have been abbreviated and represented with coloured blocks. The light purple square blocks in *Dense Blocks 1, 2, 3* and *4*, shown in Fig.4.5, are repeated multiple times. For instance, the structure of *"Dense Block 1"* is replicated six times. An *"Output 1"* shows the accuracy of the training, while *"Output 2"* represents the result of the biometric verification in terms of EERs.

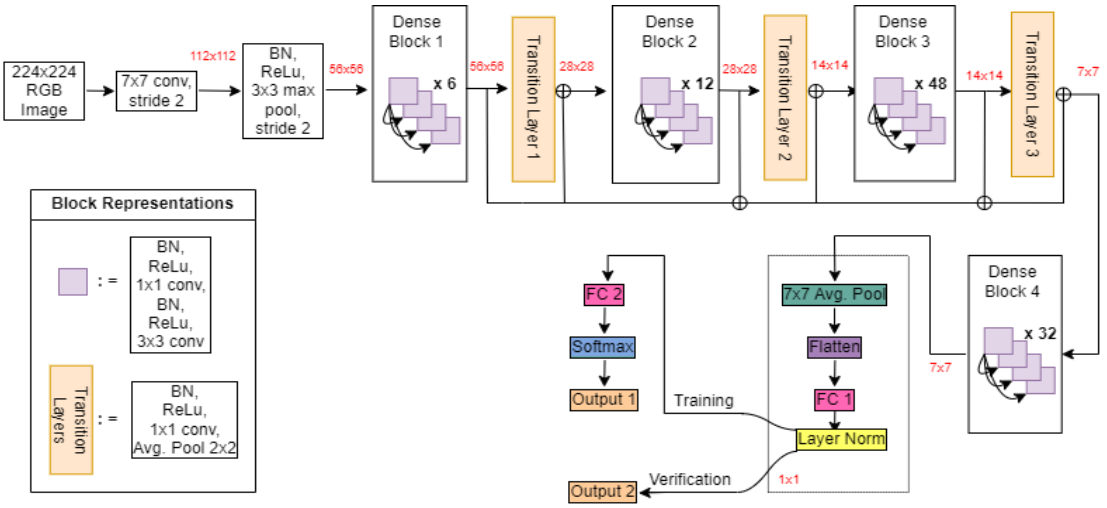

FIGURE 4.5: Used DenseNet201 model in biometric verification

The DenseNet201 architecture was trained to utilise data from 11 subjects analogous to the ResNet50 training model. The parameters used for training and the number of training and validation images (80% training / 20% validation) used are described in Section 4.2. As with the ResNet50 model, the trained model creates the *DAGNetwork*. Using this *DAGNetwork*, the samples reserved for the test are

embedded using the *"predict"* toolbox. Then, N number of samples are selected from the created embeddings and a template is created by taking the average. This process is explained in Eq.(4.1). Euclidean distances between these embeddings and the templates are computed. Distances exceeding a predetermined threshold indicated that the two samples originated from distinct subjects. Conversely, if the calculated Euclidean distance fell below the threshold, it could be inferred that the two samples belong to the same subjects.

Table 4.6 presents the mean EERs for varying numbers of genuine samples and three distinct image representations. *Sc2, Sc4,* and *Sc10* denote scalogram images generated from time windows of 2, 4, and 10 seconds respectively. *Sp2, Sp4,* and *Sp10* represent spectrogram images, while *Mel2, Mel4,* and *Mel10* correspond to Mel-spectrogram images with equivalent time windows. The designations *1-S, 3-S,* and *5-S* indicate the utilisation of 1, 3, and 5 genuine samples in enrollment.

TABLE 4.6: Mean EERs(%) are shown for different numbers of genuine samples and three image representations. *Sc2, Sc4 and Sc10*: Scalograms. *Sp2, Sp4 and Sp10*: Spectrograms. *Mel2, Mel4 and Mel10*: Mel-spectrograms in 2 seconds, 4 seconds and 10 seconds time windows. *1-S, 3-S, and 5-S*: The number of genuine samples in enrollment.

Devices	Faros			Hexoskin		
# of Genuine Samples / Image Types	1-S	3-S	5-S	1-S	3-S	5-S
Sc2	36.55%	33.58%	34.20%	35.78%	31.12%	30.79%
Sc4	22.98%	**16.85%**	**12.81%**	22.49%	**10.60%**	**8.88%**
Sc10	**22.25%**	20.74%	20.88%	20.47%	19.92%	18.13%
Sp2	39.48%	38.56%	37.35%	38.78%	36.58%	35.30%
Sp4	40.81%	36.52%	36.21%	33.95%	29.29%	27.76%
Sp10	26.65%	26.79%	26.24%	**19.58%**	19.03%	18.20%
Mel2	36.44%	36.58%	38.16%	39.26%	38.93%	41.06%
Mel4	38.97%	35.60%	35.78%	38.56%	37.21%	35.62%
Mel10	33.10%	35.16%	34.34%	29.17%	28.90%	26.85%

Despite being a consumer-based device, the Hexoskin usually yields the lowest EERs. In the Hexoskin device, scalogram images produced superior results in shorter time windows (e.g. 2 seconds and 4 seconds) compared to other time-frequency representations. Conversely, spectrogram images achieved lower EERs

within a 10 seconds time window. Although Mel-spectrograms generally exhibited the highest EERs, their performance was comparable to other representations within a 2 seconds time window for both devices. As a general trend, a decrease in EERs is observed when transitioning from 2 seconds to 10 seconds time windows. However, an exception to this trend is noted in the case of scalogram images, which yield the lowest EERs within a 4 seconds time window.

4.3 Discussion

During the first phase of our study, we examined ECG signals in biometric verification systems. These signals were obtained from various devices and underwent different data collection processes. Additionally, they encompassed a range of emotional states and physical activities.

These studies investigated the comparison between medical and wearable devices in real-life biometric verification scenarios, analysed the impact of various features on biometric verification systems in ML, and evaluated the performance of DL models (i.e. DenseNet201 and ResNet50 CNN architectures) applied to three time-frequency representations. The general outcomes of the Phase 1 is shown in Table 4.7.

TABLE 4.7: The general outcomes of the Phase 1

Studies	Min - Max EERs (%)
E-HOL (Section 4.1.1)	3.64% - 45.67%
WeSAD (10s) (Section 4.1.1)	1.60% - 48.79%
WeSAD (4s) (Section 4.1.2)	19.81% - 36.43%
Vollmer (ML) (Section 4.1.2)	17.36% - 33.51%
Vollmer (ResNet50) (Section 4.2.1)	11.15% - 39.64%
Vollmer (DenseNet201) (Section 4.2.2)	8.88% - 41.06%

In a real-life biometric verification scenario, the performance of different classifiers was evaluated at varying enrollment times. Results indicate that the EER decreases as enrollment time increases. Due to their high performance, the NB (best EERs: **3.64%** E-HOL and **3.02%** WeSAD) and DT (best EERs: **7.89%** E-HOL and **1.60%** WeSAD) classifiers were selected for use in subsequent studies. The E-HOL dataset, which was prepared using a medical device and had a larger sample size, longer total recording time, and more participants, yielded higher EER results compared to the WeSAD dataset which utilised a wearable device. It is hypothesized that unlabelled activities and varying emotional states may have had an impact on ECG recordings in the E-HOL dataset. While the 15 features used in this study were found to be distinctive for some classifiers, they exhibited lower performance in classifiers such as KNN and LDA. This observation raises the question of whether newly derived features may yield better performance.

The Manhattan and Euclidean distances are commonly employed methodologies for quantifying the dissimilarity between two templates in the context of biometric verification. However, their utilisation as features represents a novel approach. The fundamental premise is that the positions, amplitudes, and temporal intervals between the P, Q, R, S and T peaks provide sufficient information for biometric verification and identification. Consequently, the inter-peak amplitude distances may also serve as a distinguishing characteristic. In this study, the Vollmer dataset was employed in lieu of the E-HOL dataset. The primary rationale for this selection is the availability of labelled physical activities and the concurrent acquisition of data from four distinct ECG recorders within the Vollmer dataset, thereby facilitating inter-device comparisons. 4 seconds time windows were used instead of 10 seconds time windows because research in the literature showed that lower-sized time windows could also achieve higher results [178].

The sample size for both training and testing exceeds that of our studies described in Sections 4.1.1 and 4.2, providing the opportunity for future studies to be structured in accordance with real-world scenarios. Nonetheless, the selection of these rates was necessitated to ensure comparability with our studies conducted during Phase 2 (Chapter 5) and Phase 3 (Chapter 6). It has been observed that the

proposed features are more successful in physical activities than emotional states. Furthermore, the proximity of the FAR and FRR ratios suggests that the system does not exhibit imbalances such as excessive acceptance or rejection. In general, the NB classifier exhibits lower EERs compared to the DT classifier. Despite its status as a medically approved device, the SomnoTouch device yielded the highest EER results for both classifiers. Conversely, the Faros device frequently produced the lowest EER results. The Nexus-10 and Hexoskin devices demonstrated comparable EER results and were competitive with the Faros device. The findings of this study are promising for wearable technology, as the T-shirt-based Hexoskin device produced results similar to those of other medically-approved devices.

The minimum EER was achieved at the M-30% case. Specifically, at M-30%, the Faros device yielded an EER of **17.36%** utilising the NB classifier, while the Nexus-10 device produced an EER of **24.87%** utilising the DT classifier. Features extracted utilising the Manhattan distance metric provided lower EERs compared to those extracted utilising the Euclidean distance metric. Nevertheless, the minimum EERs for both feature extraction methods were achieved on the same devices. Specifically, at E-30%, the Faros device produced an EER of **19.92%** utilising the NB classifier and the Nexus-10 device yielded an EER of **25.67%** utilising the DT classifier. It has been observed that physical activity data yields better results than data reflecting emotional states. This outcome diverges from the findings of studies conducted on the E-HOL and WeSAD datasets. However, when considering the Vollmer and WeSAD datasets, which have a similar number of subjects, it can be inferred that physical activities provide lower EER results.

In investigations involving short enrollment times and novel features, it has been experimentally demonstrated that increased training time results in a decreased EER. It has been observed that the 1D DL model generally outperforms classical ML models in terms of performance at short enrollment times [1]. Comparisons were made between CNN models utilising 2D time-frequency representations and those with varying enrollment times. The primary rationale for reducing the number of samples in the creation of a template is to simulate a real-life biometric verification scenario within the context of the study. The performance of the Faros

and Hexoskin devices, which demonstrated superior performance in the classical ML model, was compared. The inability to compare other devices is attributed to the constraints imposed by the COVID-19 pandemic. Incorporating devices that were not utilised in this study into future research presents an opportunity for a more comprehensive examination of device performances.

The performance of DL models is influenced by a multitude of factors, including the structure and distribution of the data, as well as the attributes extracted from the learning structure. In the literature, there are studies in which the ResNet model outperforms the DenseNet model [238], while in others, the opposite is true [226]. In our results, the DenseNet201 structure gave lower EER results in 31 out of a total of 54 results, while the ResNet50 model gave better results in the remaining 23. In the analysis of scalogram and spectrogram images obtained from the Faros device, as well as Mel-spectrogram images obtained from the Hexoskin device, the ResNet50 model demonstrated generally better performance compared to the DenseNet201 model.

It is anticipated that an increase in enrollment times would result in a reduction of EERs. In fact, all cases, including the case of the longest enrollment time in this study (i.e. 50 seconds enrollment time), are referred to as short enrollment times in the literature [1]. The data corroborate the aforementioned hypothesis, indicating a positive correlation between a decrease in enrollment time from *5-S* to *1-S* and an increase in EERs. We achieved the best EER of **11.15%** from the Faros device in ResNet50 and **8.8%** from the Hexoskin device from DenseNet201 CNN models. If we compare our results with the one-dimensional CNN structure in studies [1, 66], it is seen that higher EERs are obtained. However, while 160 subjects were used as the CNN training set in that study, only 11 subjects were used in our study, and this difference may be due to the training set.

The best results were achieved using a 10 seconds time window for the analysis of spectrogram and Mel-spectrogram images, whereas a 4 seconds time window yielded the best results for scalogram images. Scalogram images are the images in which the heartbeats are most prominently displayed, therefore, in general, the

best results can be expected to come from these data. However, as mentioned earlier, the features extracted in DL can be quite different. Altering the time window size can increase or decrease the number of heartbeats depicted in the image. For instance, a 10 seconds time window encompasses more heartbeats and consequently alters the information learned by the system. This is a critical factor that directly impacts the system's verification success. In addition, the Hexoskin device generally gave lower EER results than the Faros device in the DenseNet201 model, while the Faros device gave lower EER results in the ResNet50 model than the Hexoskin device.

Chapter 5

Phase 2: Activity Effects of ECG Biometric Verification: Activity Classification

This chapter explores the effects of new features on activity and emotional state classifications in the classical ML model. Furthermore, the impact of three distinct time-frequency representations of ECG signals is evaluated for activity classification within the context of three DL models, each employing varying epoch sizes and optimization algorithms. Extracted features and classifiers for ML are explained in subsection 3.2.2.2 and subsection 3.3. Generated time-frequency representations for DL are explained in detail in subsection 3.2.3.

The objective of this section is to investigate the influence of activities on the EER outcomes obtained through the direct biometric verification model in Phase 1 (Chapter 4). To address this research question, the efficacy of the generated features and images in activity classification was evaluated.

Section 5.1 explains activity classification utilising classical ML models, while Section 5.2 clarifies upon the application of DL models for the same purpose.

5.1 Machine Learning Models for Activity Classification

In an effort to enhance the performance of biometric models, Phase 2 was implemented, in which activity classification was utilised prior to biometric verification. During transitions between activities, the periodicity of ECG signals changes, which can result in an increase in errors in biometric verification tasks. In Phase 2, samples undergo activity classification before being subjected to specifically calibrated biometric verification stages in Phase 3.

The performance of the extracted features within ML models with the commonly used 6 classifiers were assessed in the context of both physical activity classification and emotional state classification. In addition, ECG data obtained from wearable and medical devices were evaluated with different training and test rates. This study shows the effect of training set ratios on classification.

5.1.1 Physical Activity Classification

This study introduces novel features and evaluates their efficacy in the context of activity classification. To gain a comprehensive understanding of the performance of these features across various training and testing sizes, multiple classifiers were employed, Bagged Tree Ensemble (BT), Subspace KNN Ensemble (S-KNN), Neural Networks (NN), Decision Trees (DT), Support Vector Machines (SVM) and k-Nearest Neighbours (KNN). Activity classification was conducted using 80%, 60%, and 50% randomly selected training samples collected from Faros, SomnoTouch, Nexus-10 and Hexoskin devices. The general methodology such as pre-processing, extracted features and classifiers which are used in this experiment are explained in Chapter 3. The outline of the activity and emotion status classification is shown in Fig.5.1.

All 13 subjects in the Vollmer dataset were used in the activity classification task. The number of samples per activity for each person in each training/testing sample

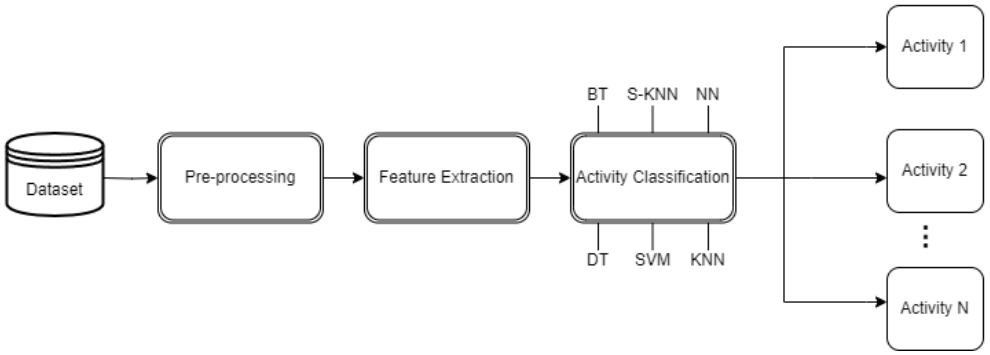

FIGURE 5.1: The outline of the activity and emotion status classification

ratio is shown in Table 5.1. The rationale for presenting the number of test samples associated with each activity per person is that, following activity classification, these samples will proceed to the biometric verification stage in Phase 3 (Chapter 6) in accordance with the specified number of test samples. As mentioned in Section 3.1.3, there are 4 physical activities in the Vollmer dataset: standing, walking, uphill walking and resting. The total number of samples is different on each device. For instance, the Faros device has 14781 (13 subjects $\times$ 1137 samples) samples for each activity.

TABLE 5.1: Training and testing sample ratios used in activity classification. The number of samples included in each activity is shown in the table per person.

Ratios (%)	**Device Names** / **Sample Sets**	**Faros**	**Somno-Touch**	**Nexus -10**	**Hexoskin**
80%	Train	900	957	916	1041
	Test	237	240	230	261
60%	Train	700	718	687	781
	Test	437	479	459	521
50%	Train	569	599	573	651
	Test	568	598	573	651

The mean of the activity classification accuracy is shown in Table 5.2 for the Vollmer dataset with resting, walking, standing and uphill walking activities. In the Table 5.2, the best results are highlighted for both feature sets. The notation *Tr-80%* denotes a training set comprising 80% of the samples, while *Tr-60%* and *Tr-50%* represent training sets containing 60% and 50% of the samples, respectively.

TABLE 5.2: Mean activity classification accuracy rates of the Vollmer dataset on different classifiers and feature sets. Tr-80% represents 80% training, Tr-60% is 60% training and Tr-50% is 50% training samples.

Vollmer Dataset		**Manhattan Distance Features**				**Euclidean Distance Features**			
Device Names		Faros	Somno-Touch	Nexus	Hexoskin	Faros	Somno-Touch	Nexus	Hexoskin
DT	**Tr-80%**	60.30%	57.93%	53.80%	57.00%	58.00%	54.68%	52.11%	56.70%
	Tr-60%	61.40%	53.80%	52.20%	62.20%	60.10%	52.18%	52.60%	58.70%
	Tr-50%	51.80%	54.81%	53.00%	60.80%	60.61%	58.76%	52.41%	58.60%
SVM	**Tr-80%**	69.20%	70.46%	60.40%	74.80%	69.20%	68.32%	61.63%	76.02%
	Tr-60%	66.20%	70.01%	58.80%	74.50%	66.14%	68.10%	59.74%	72.50%
	Tr-50%	65.50%	66.83%	60.70%	74.50%	66.12%	66.93%	59.43%	74.10%
KNN	**Tr-80%**	66.00%	70.46%	58.50%	74.40%	65.10%	70.00%	61.18%	73.10%
	Tr-60%	62.30%	71.23%	59.40%	72.90%	63.10%	69.97%	58.12%	72.80%
	Tr-50%	62.70%	67.94%	56.30%	72.00%	62.30%	67.62%	55.67%	71.30%
BT	**Tr-80%**	88.20%	80.39%	84.90%	86.70%	86.20%	80.26%	86.14%	85.70%
	Tr-60%	85.80%	77.72%	81.90%	85.60%	89.10%	79.32%	82.50%	84.50%
	Tr-50%	86.60%	77.50%	80.80%	86.30%	87.30%	78.12%	83.24%	84.40%
S-KNN	**Tr-80%**	97.05%	96.88%	**97.94%**	96.94%	**97.05%**	96.56%	96.96%	96.26%
	Tr-60%	93.42%	93.58%	93.08%	**94.77%**	93.19%	93.48%	92.87%	**94.19%**
	Tr-50%	91.77%	91.64%	91.41%	**92.78%**	91.33%	91.35%	90.79%	**92.35%**
NN	**Tr-80%**	68.80%	66.00%	58.20%	67.30%	66.70%	65.61%	62.79%	67.20%
	Tr-60%	64.50%	62.95%	57.30%	70.00%	66.10%	64.53%	57.72%	69.80%
	Tr-50%	63.30%	64.00%	57.20%	68.40%	63.67%	63.86%	55.54%	67.90%

An observable trend is that classification rates generally decline as the proportion of samples in the training set decreases from *Tr-80%* to *Tr-50%*. Comparable results were obtained for both Manhattan and Euclidean feature sets. The S-KNN ensemble classifier demonstrated superior activity classification accuracy relative to the other classifiers. Consequently, samples classified by activity using S-KNN were selected for biometric verification. For the best-performing S-KNN classifier, Manhattan distance features appear to have a higher classification percentage than Euclidean distance features.

Both Support Vector Machines (SVM) and KNN classifiers demonstrated particularly successful results when applied to the Hexoskin device. In terms of accuracy rate, the BT classifier ranks second. The most accurate results were achieved through the use of ensemble methods: BT and S-KNN. Upon examination of the performances of weak classifiers DT and KNN, it is evident that KNN outperforms DT. Consequently, it is unsurprising that S-KNN yields better results than BT.

Upon comparing the performance across devices, it was found that the Hexoskin and Faros devices exhibited superior performance, while the Nexus device consistently demonstrated the poorest performance across all classifiers. There may be many reasons for this, some of these reasons being that the features being utilised are inadequate for distinguishing the measured data from the Nexus device, weak classifiers were insufficient to classify the features, or the signal resampled from 8000 Hz to 256 Hz by Vollmer et al. [239] may have lost its distinctive features. The study [239] also indicated that Nexus device has a participant whose data is unstable for their analysis.

In addition to examining device performance, the classification performance of activities was also analysed by selecting the S-KNN classifier, which has the highest classification rates. Table 5.3 shows the performances of each activity classification from the best-performed S-KNN classifier. The best results of each device are indicated in bold.

TABLE 5.3: The performances for each activity class from the best-performed S-KNN classifier

		Manhattan Distance Features				Euclidean Distance Features			
	Activities \ Device Names	Faros	Somno-Touch	Nexus	Hexoskin	Faros	Somno-Touch	Nexus	Hexoskin
Tr-80%	**Resting:**	**97.89%**	**98.75%**	96.52%	**99.62%**	**98.31%**	**99.17%**	**98.70%**	**97.70%**
	Walking:	97.47%	93.33%	98.26%	93.87%	97.89%	94.17%	96.52%	94.64%
	Standing:	97.47%	97.92%	**99.13%**	98.47%	96.63%	96.25%	94.35%	95.40%
	Uphill:	95.37%	97.50%	97.82%	96.94%	95.36%	96.67%	98.26%	97.32%
Tr-60%	**Resting:**	**94.97%**	94.78%	**93.90%**	97.51%	**95.42%**	**98.12%**	**94.34%**	**97.70%**
	Walking:	92.68%	89.77%	92.16%	89.64%	92.68%	87.47%	93.03%	90.21%
	Standing:	94.05%	**95.62%**	92.81%	93.28%	93.82%	94.78%	91.50%	92.71%
	Uphill:	91.99%	94.16%	93.46%	**98.66%**	90.85%	93.53%	92.59%	96.12%
Tr-50%	**Resting:**	**94.01%**	**95.82%**	91.27%	**96.93%**	**95.60%**	**95.32%**	**92.67%**	**97.70%**
	Walking:	92.25%	85.62%	91.27%	86.33%	90.14%	85.12%	88.31%	84.49%
	Standing:	93.66%	94.31%	90.05%	91.86%	92.61%	94.98%	91.97%	91.86%
	Uphill:	87.15%	90.80%	**93.02%**	96.01%	86.97%	89.97%	90.23%	95.39%

According to the results obtained, the resting activity usually has the highest classification accuracy. In addition, standing activity ranks second, while walking and uphill walking activities generally have lower classification accuracy rates.

5.1.2 Emotional Status Classification

This study presents novel features and assesses their effectiveness in the field of emotion status classification. The same physical activity classification methods described in the subsection 5.1.1, including the same classifiers and training/testing ratios, were used to comprehensively evaluate the performance of these traits across varying emotion statuses. Chapter 3 details the general methodology utilised in this experiment, including pre-processing, feature extraction, and classifiers.

For the emotion status classification, all 15 subjects from the WeSAD dataset were utilised. Table 5.4 displays the number of samples per activity for each individual, according to each training/testing sample ratio.

TABLE 5.4: Training and testing sample ratios used in emotion status classification. The number of samples included in each emotion is shown in the table per person.

Ratios (%)	**Device Names** / **Sample Sets**	**RespiBAN**
80%	Train	886
	Test	221
60%	Train	664
	Test	443
50%	Train	554
	Test	553

A mean of the activity classification accuracy is shown in Table 5.5 for the WeSAD dataset with baseline, stress and amusement emotional status. Within the table, optimal results for both feature sets are emphasized. The *Tr-80%* notation signifies a training set consisting of 80% of the samples. Similarly, *Tr-60%* and *Tr-50%* denote training sets comprising 60% and 50% of the samples, respectively.

A noticeable trend indicates that classification rates decrease as the proportion of samples within the training set diminishes from *Tr-80%* to *Tr-50%*. Comparable results were obtained for both Manhattan and Euclidean feature sets, akin to physical activity classification outcomes. For example, with SVM, DT and KNN classifiers, Euclidean distance features yielded higher accuracy rates than

TABLE 5.5: Mean activity classification accuracy rates of the WeSAD dataset on different classifiers and feature sets. Tr-80% represents 80% training, Tr-60% is 60% training and Tr-50% is 50% training samples.

WeSAD dataset	RespiBAN device	Manhattan Features	Euclidean Features
DT	**Tr-80%**	58.67%	58.98%
	Tr-60%	57.64%	56.13%
	Tr-50%	53.71%	55.39%
SVM	**Tr-80%**	65.86%	73.55%
	Tr-60%	63.00%	72.62%
	Tr-50%	62.34%	65.41%
KNN	**Tr-80%**	62.74%	68.51%
	Tr-60%	60.36%	67.83%
	Tr-50%	57.03%	66.79%
BT	**Tr-80%**	92.61%	82.46%
	Tr-60%	83.75%	81.92%
	Tr-50%	78.78%	77.76%
S-KNN	**Tr-80%**	**93.67%**	**90.35%**
	Tr-60%	**86.46%**	**85.78%**
	Tr-50%	**82.10%**	**80.29%**
NN	**Tr-80%**	70.52%	71.90%
	Tr-60%	68.56%	68.03%
	Tr-50%	65.45%	64.68%

Manhattan distance features. The S-KNN ensemble classifier exhibited superior accuracy in emotional status classification compared to other classifiers. As such, samples classified by emotion using S-KNN were selected for biometric verification, mirroring the methodology employed in physical activity classification. For the top-performing S-KNN classifier, Manhattan distance features demonstrated a higher classification percentage than Euclidean distance features.

Given that Vollmer and WeSAD datasets were collected under varying circumstances, direct comparisons between them are not feasible. Nonetheless, as a general observation, despite the promising results achieved by NN, SVM, BT, and S-KNN classifiers, emotion status classification outcomes were found to be inferior to those of physical activity classification.

By analysing the performance of the S-KNN classifier, which exhibits the highest mean performance, for each emotional state, it is possible to determine which emotional states can be more accurately classified using the extracted features. Table 5.6 presents the classification performance of each emotional state using the

best-performing S-KNN classifier. The optimal results for each device are denoted in bold.

TABLE 5.6: The performances for each emotional status from the best-performed S-KNN classifier

	Features / Activities	Manhattan	Euclidean
Tr-80%	**Baseline:**	92.76%	88.23%
	Stress:	93.67%	90.95%
	Amusement:	**94.57%**	**91.86%**
Tr-60%	**Baseline:**	84.88%	84.67%
	Stress:	87.13%	84.88%
	Amusement:	**87.36%**	**87.79%**
Tr-50%	**Baseline:**	81.19%	79.10%
	Stress:	82.10%	**81.29%**
	Amusement:	**83.00%**	80.48%

According to the results obtained, the emotional state of amusement has the highest classification percentage in almost every case. However, while the feeling of stress is in second place, the lowest percentage of classification belongs to the baseline emotional status. While the exact reason for this is unclear, the classifiers and feature sets used can affect classification performance. Additionally, the ability to distinguish between more stress and amusement states than the baseline emotional state suggests that the model used is better able to determine changes in heartbeats. It was observed that the feature set utilising Manhattan distances generally achieved higher results for each emotional state. Furthermore, it was found that when more training samples were used, each emotional state exhibited a higher classification percentage.

5.2 Deep Learning Models for Activity Classification

DL models necessitate a larger quantity of training data in comparison to classical ML models. Thus, the training/testing ratios were maintained constant to ensure

an ample volume of training data. We conducted evaluations with varying time window sizes, CNN architectures, and DL hyperparameters for physical activity classification purposes. All data were separated as 72% training, 8% validation and 20% testing data randomly for all GoogleNet, ResNet50 and DenseNet201 CNNs. The number of training, testing and validation images for activity classification is shown in Table 5.7. To explain the number of images in the table, for example, for images consisting of 2-second time windows (spectrogram, scalogram or mel-spectrogram), 142 images were produced per activity for each person. According to this case, each person has 568 images and 7384 images of 13 subjects were created. Of the 7384 images, 5316 were used for training (∼72%), 592 for validation (∼8%) and 1476 (∼20%) for testing.

TABLE 5.7: 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

In this study, the training process was optimized 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 32, specified the subset of the training set used in each iteration. These parameters have not been changed to be compatible with Phase 1 and Phase 3. 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 optimizers, adam (Adaptive Moment Estimation) and sgdm (Stochastic Gradient Descent with Momentum), were compared. These optimizers 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 [240].

Sgdm is an advanced version of Stochastic Gradient Descent (SGD) which utilises momentum to enhance the training process. Through the use of exponential moving averages of previous gradients, momentum supports the optimizer in navigating local minima and saddle points [241]. Conversely, adam is an adaptive learning rate optimizer that calculates individual learning rates for distinct parameters. Adam offers an efficient and robust optimization method using two other popular optimization algorithms (i.e. AdaGrad and RMSProp) [242]. The choice between sgdm and adam depends on the particular problem and architecture being utilised. While some studies have demonstrated that sgdm exhibits superior generalization compared to adam, others have found that adam converges more rapidly [243].

The activity classification performance was evaluated using the wearable Hexoskin device and the medical Faros device. Comparisons were made based on different signal representations, time windows, optimizer types, the number of epochs and three different CNN models. The proposed CNN models are explained in the following subsections for activity classification tasks.

5.2.1 GoogleNet CNN

A GoogleNet, also known as Inception, comprises 9 inception layers and is commonly employed to preserve fine details within images. The architecture aims to achieve high accuracy while minimizing computational costs compared to previous CNN models. The GoogleNet utilises filter sizes of 5x5, 3x3, and 1x1 to partition images of varying resolutions, thereby capturing more information from the image and addressing the issue of redundant information. Fig.5.2 illustrates the fundamental structure of GoogleNet in activity classification.

In the 2022a version of MatLab, the GoogleNet architecture, which consists of 22 layers and has been pre-trained on the ImageNet dataset to classify 1000 classes, is available to users. In the context of neural networks, layer freezing refers to

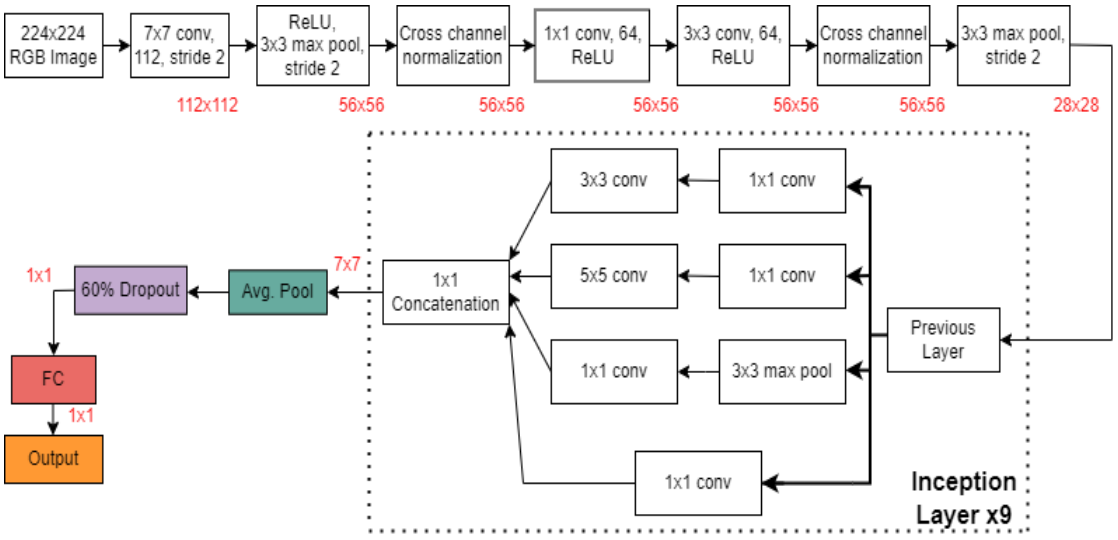

FIGURE 5.2: GoogleNet model in activity classification

the process of constraining specific layers' weights to remain non-trainable while permitting updates to other layers during training. This technique is commonly employed in transfer learning scenarios, where pre-trained models are fine-tuned for specific tasks by selectively freezing and unfreezing layers based on their relevance to the new task [244]. Layer freezing were used to adjust pre-trained model for our task. We freezed the layers of GoogleNet except the last three layers (final layers). Then these final layers indicated in colour have been incorporated for the purpose of 4 classes classification. The remaining layers are present in the original GoogleNet architecture.

A *dropout* is a regularization technique employed to mitigate over-fitting in NN, achieved by randomly omitting neurons during the training phase. Consequently, the omitted neurons neither partake in the forward pass nor in backpropagation [245]. Furthermore, due to the significant computational expenses involved, it was necessary to include a dropout layer. Despite implementing a 60% dropout layer, the training process still takes a considerable amount of time. 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 CNN model for activity classification is shown in Table 5.8. The highest accuracy rates for each device are denoted in bold.

TABLE 5.8: GoogleNet CNN accuracy rates(%) in activity classification.

	Faros Device				Hexoskin Device			
	10 ADAM	**20 ADAM**	**10 SGDM**	**20 SGDM**	**10 ADAM**	**20 ADAM**	**10 SGDM**	**20 SGDM**
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%**

It is generally observed that the results obtained are lower than the results obtained from classical ML. In this regard, if we evaluate the possible reasons according to the types of images, although the highest results are obtained in scalogram images, the overall performance is low. The possible reason for this may be that the ImageNet dataset containing natural image classification does not work well in medical image tasks [246, 247]. For spectrogram and Mel-spectrogram images, which have even lower performance, insufficient quantity data and non-informative parts of images may be shown in addition to mismatch with the task.

Upon analysis of the results, it was observed that the Hexoskin device obtained higher accuracy rates for spectrogram images in comparison to the Faros device. Nonetheless, the Faros device demonstrated superior performance with respect to scalogram and Mel-spectrogram images.

The rationale for not exceeding 20 epochs during training is that, despite the considerable time investment, there is minimal improvement in training accuracy, validation accuracy, and test results. In certain instances, even 10 epochs yielded superior accuracy compared to 20 epochs. However, utilising 20 epochs yields superior results compared to 10 epochs in general.

Experimental evidence suggests that both optimization algorithms exhibit comparable performance when applied to the GoogleNet CNN architecture. With the Faros device, 10 epochs of sgdm typically yield superior accuracy rates compared to 10 epochs of adam, whereas the inverse is true when utilising 20 epochs. In

the case of the Hexoskin device, when employing spectrogram and scalogram images, 10 epochs of adam outperforms 10 epochs of sgdm. However, when using Mel-spectrogram images, 10 epochs of sgdm achieves significantly higher results. With 20 epochs, sgdm attains slightly higher results compared to adam. Notably, in the case of Mel-spectrogram images, the adam optimizer exhibited markedly low accuracy rates.

5.2.2 ResNet50 CNN

The original ResNet50 structure described in Section 4.2.1 was preserved until the end of Stage 4 in the activity classification task. Rather than training the model from scratch, a pre-existing model, pre-trained on the extensive ImageNet dataset, was utilised (i.e. instead of creating a *DAGNetwork* ourselves, we adapted a previously created *DAGNetwork* to our system). We applied layer freezing, similar to the GoogleNet model, by freezing layers up to Stage 5. We then customised the layers in the final stage to align with our specific classification task. This approach was adopted due to the vast quantity of images contained within the ImageNet dataset and the pre-trained ResNet50 model is already capable of classifying 1000 distinct categories. In Stage 5, the *New FC* layer is restructured to adapt the pre-trained ResNet50 model, originally designed to classify 1000 categories, to classify 4 classes. Fig.5.3 illustrates the fundamental structure of ResNet50 in activity classification.

The four activities that were classified included resting, walking, standing and uphill walking on the treadmill. Subsequently, the mean accuracy rate for all activities was computed based on the *Output* layer. A mean accuracy rate was obtained from the ResNet50 CNN model for activity classification is shown in Table 5.9. The maximum accuracy rates for each device are indicated in bold.

In DL studies, similar datasets are selected to ensure that the tasks match. For example, for object classification, the model is trained with a dataset containing many objects' images such as ImageNet [248], while for activity classification, the

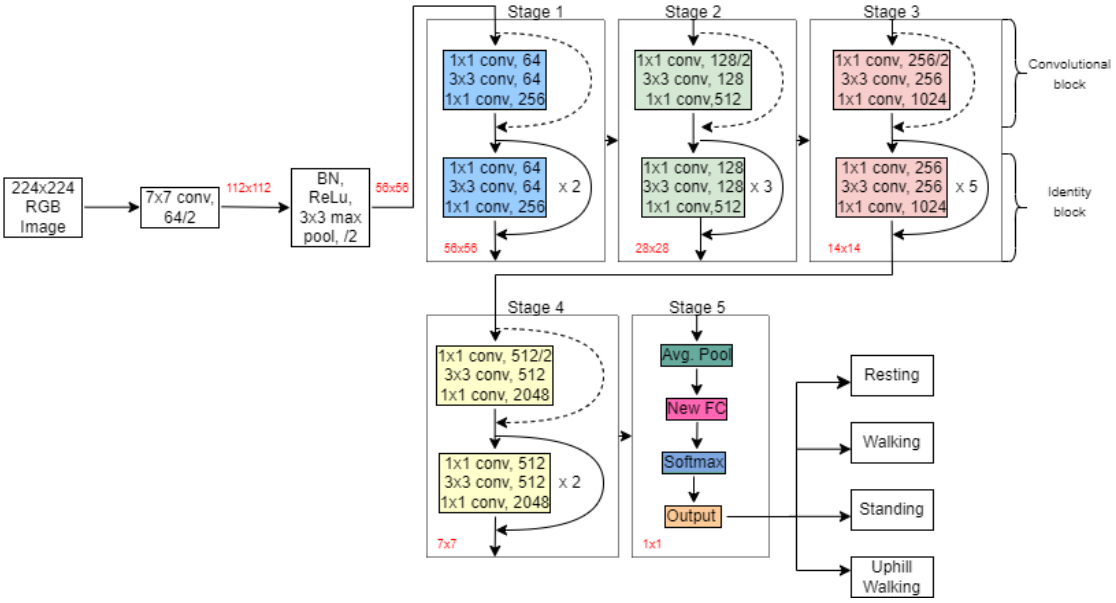

FIGURE 5.3: ResNet50 model in activity classification

TABLE 5.9: ResNet50 CNN accuracy rates (%) in activity classification.

	Faros Device				Hexoskin Device			
	10 ADAM	**20 ADAM**	**10 SGDM**	**20 SGDM**	**10 ADAM**	**20 ADAM**	**10 SGDM**	**20 SGDM**
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%

model should be trained with a dataset suitable for this purpose [247]. In cases where the task and the dataset do not match, large datasets suitable for the desired task are selected and the model is trained and then tested with the target dataset [1, 249]. For instance, Bıçakcı et al. [1] trained the 1-D DL model with E-HOL dataset and tested with WeSAD dataset. In addition, Sarkar et al. [108] used 4 different datasets which consist of different emotional states in CNN model. They trained the model with all datasets and they used 6 different types of transformed signals to feed the model. However, this method was not preferred for our CNN models due to the small number of datasets consisting only of ECG data and physical activity. Although the results were not successful in general, it achieved

more successful results than the GoogleNet model especially for mel-spectrogram image cases. Results can be improved with different parameters and fine-tuning techniques [250]. However, due to time constraints and the desire to maintain consistency with Phases 1 and 3, alternative parameters were not chosen.

In broad terms, the Faros device demonstrated superior classification outcomes compared to the Hexoskin device. It was observed that models with 20 epochs and sgdm optimizers in 10 seconds time windows yielded more successful results than those with 10 epochs and adam optimizers. With respect to the Faros device, the adam optimizer typically obtains higher accuracy rates compared to the sgdm optimizer, while the inverse is true for the Hexoskin device. Consequently, it cannot be concluded that either the adam or sgdm optimizers possess a distinct advantage over another based on these findings.

Although 20 epochs generally result in higher accuracy rates than 10 epochs, using scalogram images with 10 adam produced significantly better accuracy rates, especially when using the Faros device. Furthermore, the maximum accuracy rate among all ResNet50 models is attributed to the 10 adam scenario, in which scalogram images with a 4 second time window are employed.

5.2.3 DenseNet201 CNN

The original DenseNet201 structure described in Section 4.2.2 was preserved until the end of *Dense Block 4* in the activity classification task. Similar to GoogleNet and ResNet50, the *DAGNetwork*, pre-trained with the ImageNet dataset, was modified to accommodate our specific 4-class classification study. In this technique, we start by training a base network using the ImageNet database. Next, we transfer the convolutional filter layers from this base network to the target network [250]. The remaining fully connected layer in the target network is started with random initialization. Finally, we freeze the weights for all layers except the final fully connected layer, which is the only layer that undergoes training. The newly established fully connected layer (i.e. *New FC*) is configured and utilised with 4 classes

to classify the activities of resting, walking, standing, and uphill walking. The *Output* layer employed for classification purposes yields the accuracy rate of the classification. Fig.5.4 shows the main structure of DenseNet201 in activity classification. In the figure, the internal structure of the shortened coloured blocks is also indicated. In addition, the dimensions of the output procured at the conclusion of each block are denoted as "$n \times n$".

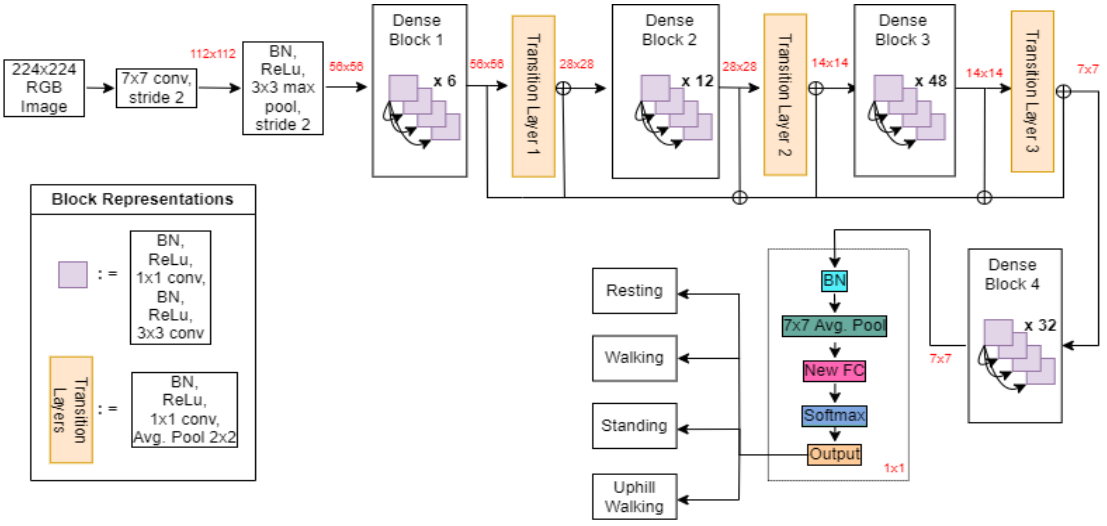

FIGURE 5.4: DenseNet201 model in activity classification

The mean execution time of training was greater than that of other ResNet50 and GoogleNet models. One possible explanation for this phenomenon could be the higher number of layers compared to other models. Additionally, as the training set contains fewer images (i.e. from the 2 second time windows case to the 10 second time windows case), the training time also decreases proportionally.

A mean accuracy rate was obtained from the DenseNet201 CNN model for activity classification is shown in Table 5.10. The highest accuracy rates are shown in bold for each device.

Upon examination of the table, it is evident that the Faros device demonstrated higher accuracy rates compared to the Hexoskin device in nearly all instances. While this observation is consistent with the ResNet50 model, it diverges in the case of the GoogleNet model.

TABLE 5.10: DenseNet201 CNN accuracy rates (%) in activity classification.

	Faros Device				Hexoskin Device			
	10 ADAM	**20 ADAM**	**10 SGDM**	**20 SGDM**	**10 ADAM**	**20 ADAM**	**10 SGDM**	**20 SGDM**
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%**

In the Faros device, the highest accuracy rates for scalogram and spectrogram images were achieved using 20 adam, while 20 sgdm yielded the highest results for Mel-spectrogram images. In the Hexoskin device, it cannot be definitively stated that 20 epochs have better results compared to 10 epochs, or vice versa. Similarly, the performances of the optimizers do not appear to be significantly better than one another. The adam optimizer demonstrated superior performance on spectrogram images, while the sgdm optimizer generally exhibited higher performance on Mel-spectrogram and scalogram images for the Hexoskin device.

5.3 Discussion

In the second phase of our study, we conducted an analysis of ECG signals within activity classification systems. These signals were derived from a variety of devices and were subjected to distinct data collection methodologies in our ML studies while being obtained through a uniform data collection process in our DL studies. Furthermore, the signals represented a diverse range of emotional and physical states.

Our investigations included a comparative analysis of medical and wearable devices in activity classification tasks, an examination of the influence of various newly created features and different training/testing sample sizes on ML systems.

In addition, the studies included an evaluation of the performance of DL architectures such as GoogleNet, DenseNet201, and ResNet50 CNNs applied to scalogram, spectrogram, and Mel-spectrogram images. In fact, the purpose of the studies in Phase 2 is to examine whether the performance of biometric verification studies in Phase 1 can be improved by activity classification. In this respect, it is the first study in the literature. The general outcomes of the Phase 2 is shown in Table 5.11.

TABLE 5.11: The general outcomes of activity/emotional states classification in the Phase 2

Studies	Min - Max Accuracy Rates (%)
Vollmer (ML) (Section 5.1.1)	51.80% - 97.94%
WeSAD (4s) (Section 5.1.2)	53.71% - 93.67%
Vollmer GoogleNet (Section 5.2.1)	25.00% - 70.81%
Vollmer ResNet (Section 5.2.2)	44.70% - 68.55%
Vollmer DenseNet201 (Section 5.2.3)	43.94% - 69.85%

Based on the ML model analysis, it appears that physical activities have a higher accuracy rate than emotional states with all features used. Additionally, increasing the number of training samples generally improved the accuracy percentage for both datasets.

To better understand our findings in relation to other research, it's important to compare and contextualize them with similar studies. Several studies have been conducted on activity recognition and classification, utilising both nonlinear features [251] such as Wavelet Transform (WT) and linear features [252] such as skewness, kurtosis, and mean of maxima and minima, applied to various datasets. Additionally, several studies have compared deep fusion [107], CNN [108], convolutional+transformer architectures [109], CNN+LSTM [110] and classical ML classifiers [105, 106] for emotion classification using the WeSAD dataset. While Studies [105, 251, 252], and [107] employed a combination of sensor data including

EMG, accelerometer, and ECG, along with features such as mean, standard deviation, peak detection, and HRV, our proposed Manhattan and Euclidean distance-based features demonstrated superior activity classification rates. In study [108], which gave slightly better results than our ML study (i.e. obtained 93.67% accuracy rate), a 95% accuracy rate was achieved in emotional states classification. Although the WeSAD dataset was used in this study, the CNN model was trained with 4 different datasets.

According to the results obtained from the DL models, the accuracy rate was highest for DenseNet201, followed by ResNet50 and GoogleNet. While 10 epochs yielded satisfactory results, 20 epoch cases were generally more successful. The sgdm optimizer produced higher results for the Hexoskin device, whereas the adam optimizer achieved higher accuracy rates for the Faros device. In the DenseNet201 model, the highest accuracy rate was obtained from 2 seconds time windows, followed by 4 seconds time windows and 10 seconds time windows. However, in the ResNet50 and GoogleNet models, the order was 4 seconds, 2 seconds and 10 seconds time windows. For all the CNN models used, the highest measured accuracy rate was obtained from the *Sc4* case in the Faros device.

The images of the scalogram, spectrogram, and Mel-spectrogram show the heartbeat, which varies in quantities depending on the time window. For instance, in the 2 seconds time window cases, there is an average of 2 heartbeats, while in the 10 seconds time window cases, there is an average of 10 heartbeats. The number of images obtained also varies according to the size of the time windows, as illustrated in Table 5.7. This shows us that more images are obtained in the 2 seconds case. Therefore, 10 seconds 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 seconds time window cases. This shows that the number of images in the training set for activity classification affects the system more than time windows. In addition, the high accuracy rates of the 4 seconds time windows indicate that both the number of images 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 utilising the same dataset, our findings provide valuable insights into the comparative effectiveness of various time-frequency representations, time windows, epoch numbers, optimizers, and CNN models. Our study represents the first application of the Vollmer dataset [3] to activity classification. Blasing et al. [239] previously employed this dataset to accurately identify Q, R, and S peaks, reporting that the Faros device exhibited the highest measurement accuracy. Although different datasets were used, it was observed that we obtained lower accuracy rates when comparing CNN models with the literature [108, 110, 253, 254]. However, although topics such as classifying arrhythmia using ECG data, classifying heart beats, classifying patients, etc. are frequently found in the literature, classifying activity with only ECG signal is an uncommon issue. In addition, a lot of training data needs to be used to train the CNN model. For this reason, efforts have been made to solve the need for large amounts of training data by combining various datasets to train the model [108], create an ensemble CNN model [253], or combine data obtained from different sensors [108, 254]. Using only ECG and PPG data, Almanifi et al. [253] trained Resnet50V2, MobileNetV2 and Xception CNN models, which were previously trained using ImageNet, by transfer learning. They combined the inferences obtained from each CNN model using the Concatenate layer and performed ECG-based activity classification using a new FC layer. They called this method as ensemble CNN model and they used it to improve classification accuracies obtained by only ECG signals. When these methods were examined, it was observed that both time and computational costs were high. In addition, since the use of different methods would disrupt the consistency between phases (i.e. Phases 1 and 3), these methods were not used.

In this chapter, the performances of novel features on different devices, under different training and test sample rate conditions, and different classifiers were examined by activity classification. When discussing general inferences about this chapter, it has been shown that, unlike conventional time-interval features, features based only on the distance between the amplitudes of the peaks can also

be used. In addition, considering the device performances, although medical devices are thought to obtain more reliable signals than wearable devices for various reasons (e.g. number of leads, location of the sensor, sensor type, etc.), it has been observed that the Hexoskin device, which is in T-shirt form, achieved better results than medical devices such as SomnoTouch and Nexus. When we consider the DL models, although the Faros device, which is a medical device, achieves better results than the Hexoskin device, the general classification results are not as satisfactory as ML. Considering the performances of different classifiers, it has been observed that more complex classifiers (such as ensemble models and SVM) give better results than weak classifiers (such as DT and KNN) when the number of features is not large. Considering the extent to which these results are affected by sample size, a larger amount of training data yields a higher accuracy rate, as is a general trend in classical ML models. In line with this, using a larger amount of training images in DL models increased the accuracy rate. For instance, the number of time-frequency images created from 2-second time windows is greater than the images created from 10-second time windows, so in the 2-second case, the models were trained with more images.

Chapter 6

Phase 3: Activity Effects of ECG Biometrics: Verification Following Activity Classification

In this Chapter, the questions of does activity or emotion classification prior to biometric verification improve performance in medical and wearable ECG devices and do different machine learning models and their parameters contribute to the improvement of device performances are explored. The impact of various ML models and their respective parameters on device performances were thoroughly analysed with different aspects in Chapters 4 and 5. However, in Phase 3 (Chapter 6), we directly evaluated device performance through biometric verification for each individual activity.

In Chapter 5, ECG data were categorised based on their corresponding activities using both classical ML and DL models. Subsequently, the biometric verification tasks for each class were conducted in Phase 3, and their performances were evaluated. This allowed for a comparison between the performances of medically approved ECG recorders and wearable ECG recorders. To determine the impact of activity classification on device performance during verification tasks, classical ML

models are examined in Section 6.1, while DL models are investigated in Section 6.2.

6.1 Machine Learning Models for Activity-aware Biometric Verification

In this section, we evaluate the biometric verification performance of data classified based on distinct physical activities and emotional states, utilising DT and NB classifiers for each activity and dataset. To achieve this objective, the activity classification framework depicted in Fig.5.1, as outlined in Chapter 5, has been revised. The proposed activity-aware biometric verification model is illustrated in Fig.6.1. N represents the number of activity classes in the figure. The rationale for selecting NB and DT classifiers for biometric verification in Phase 3 studies is to monitor the progression of the direct biometric verification studies described in Phase 1.

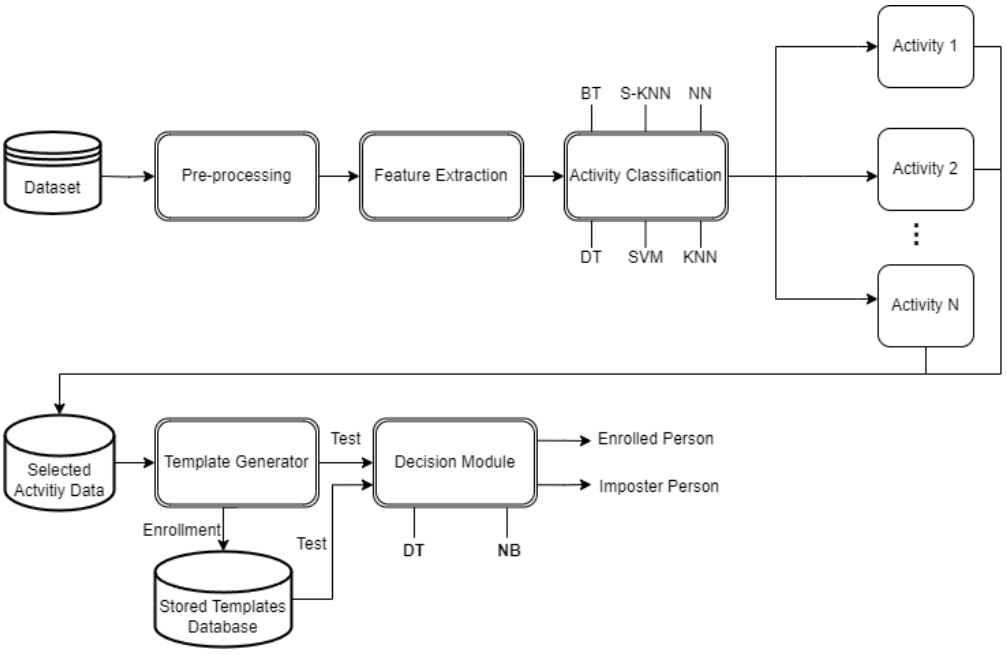

FIGURE 6.1: The proposed activity-aware biometric verification system.

The primary objective of the proposed system is to mitigate the effects of signal fluctuation that arise when transitioning between different activities. In the

proposed model, samples are initially classified based on their activity and subsequently directed to tuned biometric verification tasks, irrespective of whether the activity is correctly or incorrectly classified. While inaccurate activity classification may impact biometric verification accuracy, utilising tuned biometric verification modules according to activity may enhance overall system performance.

Removing incorrectly classified samples will not be a realistic verification scenario. Therefore, samples that were incorrectly classified during the activity classification task were evaluated in the biometric verification task using the new label assigned by the classifier of the activity classification. For instance, a sample originally labelled as '*Resting*' may be classified as '*Walking*' during the classification process. This sample will then participate in Phase 3 biometric verification studies alongside other samples in the '*Walking*' class.

It was observed that the various training/testing sample sizes (i.e. *Tr-80%, Tr-60%* and *Tr-50%* cases) examined in activity classification yielded different accuracy rates. For this reason, the number of misclassified samples and the total number of samples passed to Phase 3 differ for each case. In this context, the number of total samples examined in Phase 3 was similar to the number of test samples specified in Table 5.1. For example, for the Faros device, 237 samples per activity from each person in the *Tr-80%* case were classified according to their activity.That is, the total number of data per activity used in Phase 3 is 237 samples for each person in the *Tr-80%* case. To investigate the effects of this situation, biometric verification tasks were examined separately for *Tr-80%, Tr-60%* and *Tr-50%* cases.

In subsection 6.1.1, we analyse ECG data collected during physical activities using Faros, Somnotouch, Nexus-10, and Hexoskin devices. In subsection 6.1.2, we examine ECG data obtained with the RespiBAN device under varying emotional states. During the analysis of these data, we employed features based on Manhattan and Euclidean distances, as described in previous sections, and compared the performance of these features for each case.

6.1.1 Physical Activity-aware Models

This section focuses on the Vollmer dataset mentioned in Chapter 3, which includes recordings of ECGs from 4 devices. The performance of biometric verification for each device, namely the Faros, SomnoTouch, Nexus and Hexoskin devices, is evaluated individually in subsections 6.1.1.1, 6.1.1.2, 6.1.1.3 and 6.1.1.4 respectively.

6.1.1.1 Faros Device

During the biometric verification task, we tested the Faros device, which has received medical approval (i.e. having CE and FDA certificates) [163], using NB and DT classifiers. Additionally, the *Tr-80%, Tr-60%* and *Tr-50%* values shown in Table 6.1 are the ratios of the samples classified in the activity classification. For the biometric verification phase, we selected data from three unseen subjects similar to Phase 1. However, since the results of activity classification included incorrectly classified samples, the samples for each activity class were different. As explained in Section 6.1, samples misclassified during activity classification were not removed from the model, but instead were examined in Phase 3 with the new class label. For instance, if we examine the *Tr-80%* case of the Faros device seen in Table 5.1, 237 samples were passed for each activity from Phase 2 to Phase 3 per person. If we were to classify activities with 100% accuracy, there would be 237 samples belonging to each activity class in Phase 3, but our classification accuracy rate is 97.05% (see Table 5.2). Therefore, the number of samples for each class is different (the number of sample for this case: resting=241, walking=235, standing=242 and uphill walking=230) and this changes at each activity classification phase. Thus, we couldn't provide an exact number of samples for each class. Nevertheless, we selected 50% of the samples as genuine samples during the testing phase for the verification tasks so that the training/testing ratios could be equal, even though the sample numbers weren't equal. Table 6.1 shows the biometric verification performances in terms of EERs for both Manhattan and Euclidean distances-based features in each activity and classifier case. The best EER results are indicated in bold.

TABLE 6.1: Biometric verification performances of Faros device in terms of EER. 50% of genuine samples were used in testing.

			DT		NB	
	Feature Names		Manhattan	Euclidean	Manhattan	Euclidean
Activity Names	Resting	Tr-80%	13.37%	13.36%	6.93%	7.38%
		Tr-60%	12.25%	12.98%	**6.31%**	6.75%
		Tr-50%	11.59%	**11.27%**	6.59%	6.89%
	Walking	Tr-80%	13.78%	13.75%	7.35%	7.53%
		Tr-60%	13.47%	13.64%	7.23%	7.67%
		Tr-50%	12.72%	**12.68%**	6.63%	**6.61%**
	Standing	Tr-80%	13.36%	13.75%	7.83%	7.30%
		Tr-60%	13.31%	13.77%	7.35%	7.10%
		Tr-50%	**12.31%**	12.68%	**6.47%**	6.68%
	Uphill Walking	Tr-80%	13.43%	13.47%	7.69%	7.85%
		Tr-60%	**11.99%**	12.09%	6.84%	7.76%
		Tr-50%	12.07%	13.89%	**6.37%**	7.51%

Based on the obtained results, it can be seen that NB outperforms DT with lower EER values. In general, the best performances, as indicated by the lowest EERs, were observed in the *Tr-50%* cases, while the highest EERs were observed in the *Tr-80%* cases. The most favourable results were obtained from the NB classifier, with an EER of 6.31%, and from the DT classifier, with an EER of 11.27%, in the resting activity for both cases. In the context of interpreting results, it is important to note that the terms *Tr-80%*, *Tr-60%*, and *Tr-50%* originated from Phase 2. Only the test samples specified in Table 5.1 advanced to Phase 3. For the Faros device, Phase 3 involved examining 237 samples in the *Tr-80%* case, 437 samples in the *Tr-60%* case, and 568 samples in the *Tr-50%* case per person and activity. Stating that the randomly selected genuine and imposter samples vary in each case and are chosen in equal numbers, it can be deduced that the total number of samples per subject is higher in the *Tr-50%* case compared to the *Tr-80%* case. Consequently, when 50% of these samples are utilised as genuine samples during testing in the *Tr-50%* case, a longer enrollment time is obtained. In this context, it was observed that a larger number of enrollment samples resulted in lower overall EERs.

When we analyse the EERs of this device in terms of direct biometric verification

tasks during Phase 1, we notice that the EER results are significantly lower in Phase 3 for both the DT and NB classifiers. Even in the most favourable scenarios of M-30% (i.e. 17.36% EER for NB, 25.16% EER for DT) and E-30% (i.e. 19.92% EER for NB, 27.93% EER for DT) in Phase 1, the EERs in Phase 3 are lower than those of Phase 1. The probable reason for this is that both the training samples and the testing samples were randomly selected across activities in Phase 1. The differences between the activities negatively affected the biometric verification performance. It is observed that this device achieves the lowest EER results when compared to other devices in Phase 3.

Upon individual examination of each activity, it was observed that the Euclidean features exhibited lower EERs in the DT classifier for both *resting* and *walking* activities. Conversely, the Manhattan features demonstrated lower EER values in the NB classifier for these activities. However, an inverse trend was observed in the *standing* activity. In the *uphill walking* activity, the Manhattan features yielded lower EER values for both classifiers. While the most favourable results were observed in the *resting* activity, the *uphill walking* activity also yielded highly successful outcomes. It can be inferred that as the intensity of the exercise increased, there may have been an increase in distortions in the obtained ECG signals. It can be said that despite an increase in HR, there was also an increase in the distinctiveness of features.

6.1.1.2 SomnoTouch Device

During the biometric verification task, the medical SomnoTouch device was tested using NB and DT classifiers. The biometric verification processes were identical to those of the Faros device, as described in subsection 6.1.1.1. Table 6.2 presents the biometric verification performances in terms of EERs for both Manhattan and Euclidean distance-based features across each activity and classifier case. The most favourable EER results are indicated in bold.

In all cases, the NB classifier yielded lower EERs compared to the DT classifier. Furthermore, upon examination of the best performances for each activity, it was

TABLE 6.2: Biometric verification performances of SomnoTouch device in terms of EER. 50% of genuine samples were used in testing.

			DT		NB	
	Feature Names		**Manhattan**	**Euclidean**	**Manhattan**	**Euclidean**
Activity Names	**Resting**	**Tr-80%**	13.91%	13.89%	7.15%	7.09%
		Tr-60%	13.71%	13.37%	7.04%	7.21%
		Tr-50%	**12.17%**	12.78%	**6.67%**	7.18%
	Walking	**Tr-80%**	13.89%	13.57%	7.73%	7.74%
		Tr-60%	13.59%	12.95%	7.43%	7.17%
		Tr-50%	**12.67%**	**12.67%**	6.81%	**6.73%**
	Standing	**Tr-80%**	13.50%	14.09%	7.46%	7.92%
		Tr-60%	13.78%	13.54%	7.74%	7.42%
		Tr-50%	**12.46%**	12.86%	**7.12%**	7.85%
	Uphill Walking	**Tr-80%**	13.91%	14.24%	7.84%	8.09%
		Tr-60%	12.67%	13.66%	7.47%	7.75%
		Tr-50%	**12.15%**	12.83%	7.31%	**7.23%**

observed that the lowest EERs were obtained in the *Tr-50%* case. As in Phase 1, more enrollment samples achieved lower EER results. The results obtained for the DT classifier are almost half of those obtained in Phase 1 and for the NB classifier are almost one-third of those in Phase 1. This shows the accuracy of the proposed framework. In addition, when the performance of this device is examined in general, it is observed that it is lower than the Faros device.

Upon examination of the best performances for each classifier, it was observed that Manhattan features yielded superior results compared to Euclidean features in the DT classifier. Conversely, in the NB classifier, Manhattan features demonstrated successful outcomes in the *resting* and *standing* activities, while Euclidean features were more successful in the *walking* and *uphill walking* activities. However, when considered in a broader context, it was observed that Manhattan features generally outperformed Euclidean features. The best performances are obtained from *uphill walking* activity with an EER of 12.15% for the DT classifier, while from *resting* activity with an EER of 6.67% for the NB classifier.

6.1.1.3 Nexus-10 Device

The Nexus-10 device, which possesses both a medical CE certificate and an FDA approval certificate [162], was utilised in biometric verification models employing various classifiers, analogous to the Faros and SomnoTouch devices. All processes and parameters utilised in biometric verification are identical to those employed in the Faros and SomnoTouch devices. The EERs of biometric verification based on Manhattan and Euclidean distance features are shown in Table 6.3 for different activities and classifiers. The lowest EER values are highlighted in bold.

TABLE 6.3: Biometric verification performances of Nexus-10 device in terms of EER. 50% of genuine samples were used in testing.

			DT		NB	
	Feature Names		Manhattan	Euclidean	Manhattan	Euclidean
Activity Names	Resting	Tr-80%	13.79%	13.78%	6.82%	7.65%
		Tr-60%	12.57%	13.72%	**6.65%**	7.76%
		Tr-50%	**11.53%**	12.76%	6.72%	7.30%
	Walking	Tr-80%	13.76%	13.25%	7.41%	7.59%
		Tr-60%	13.64%	13.37%	6.91%	7.60%
		Tr-50%	**12.50%**	12.96%	**6.74%**	6.96%
	Standing	Tr-80%	13.29%	13.47%	7.59%	7.53%
		Tr-60%	12.94%	13.83%	7.18%	7.59%
		Tr-50%	**12.34%**	12.67%	**6.36%**	7.18%
	Uphill Walking	Tr-80%	13.62%	13.78%	7.57%	7.82%
		Tr-60%	12.53%	13.89%	7.39%	7.92%
		Tr-50%	**12.37%**	13.47%	**7.24%**	7.30%

The NB classifier outperformed the DT classifier across different feature sets, indicating its robustness and suitability for the task. The features derived from the Manhattan distance consistently yielded lower EERs than those based on the Euclidean distance, suggesting that the Manhattan distance-based features capture more discriminative information for biometric verification. Moreover, the EERs tended to decrease as the proportion of testing data was increased from *Tr-80%* cases to *Tr-50%* cases, implying that the classifiers were able to generalize well with more data. The best results were obtained with 11.53% EER from the DT

classifier during *resting* activity, and 6.36% EER from the NB classifier during *standing* activity.

If we look at the EER results in Phase 1, it is observed that it is much higher than Phase 3. Furthermore, the Nexus-10 device, which yielded the second most favourable results among all devices in Phase 1, maintained its position as the second most successful device in Phase 3. The original sampling frequency of this device is approximately 8000 Hz. As the sampling frequency increases, the noise in the obtained signal may increase. Although the signal preprocessing and sampling frequency is reduced, the distortions in the signal still may not be completely eliminated. This may be one of the reasons why this device is not the most successful device.

6.1.1.4 Hexoskin Device

The performance of Hexoskin, a wearable smart device [165], has been evaluated through the utilisation of biometric verification models delineated in subsection 6.1.1.1, in order to compare with other devices. The biometric verification performance of this device in terms of EER is shown in Table 6.4. In addition, the best performances in the Table are indicated in bold for each activity.

Consistent with the previous three devices, the NB classifier gave the most favourable results for this device. However, in contrast to other devices, the *Tr-60%* cases achieved highly successful outcomes similar to those of *Tr-50%* cases. Within this framework, the lowest EERs for each activity, particularly in the NB classifier, were mainly derived from *Tr-60%* cases. Similar to other devices, features based on Manhattan distances outperformed those based on Euclidean features.

Consistent with the Nexus-10, optimal performances were obtained from standing and resting activities. Within this framework, the NB classifier yielded the lowest EER of 6.31% for *standing* activity, while the DT classifier produced the lowest EER of 12.66% for *resting* activity. The minimum EER value of 6.31% attained from the Hexoskin device is the same as the minimum EER value derived from the

TABLE 6.4: Biometric verification performances of Hexoskin device in terms of EER. 50% of genuine samples were used in testing.

			DT		NB	
	Feature Names		**Manhattan**	**Euclidean**	**Manhattan**	**Euclidean**
Activity Names	**Resting**	**Tr-80%**	13.63%	13.82%	7.09%	7.67%
		Tr-60%	12.70%	13.95%	7.11%	7.53%
		Tr-50%	**12.66%**	13.62%	**6.63%**	6.91%
	Walking	**Tr-80%**	13.99%	13.66%	7.57%	8.07%
		Tr-60%	13.85%	13.64%	**6.41%**	7.98%
		Tr-50%	**12.89%**	13.07%	6.62%	7.64%
	Standing	**Tr-80%**	13.71%	14.02%	7.17%	7.78%
		Tr-60%	13.83%	14.38%	**6.31%**	7.80%
		Tr-50%	**12.82%**	13.84%	6.46%	7.30%
	Uphill Walking	**Tr-80%**	13.79%	14.18%	7.44%	7.97%
		Tr-60%	**12.90%**	14.06%	**6.46%**	7.91%
		Tr-50%	13.01%	13.76%	6.58%	7.49%

Faros device and outperforms the minimum performance of both the SomnoTouch and Nexus devices utilising the NB classifier. Conversely, the 12.66% EER value derived from the Hexoskin device utilising the DT classifier exceeded the minimum values of the other three devices employing the same classifier, indicating insufficient performance.

The dissimilarity in the outcomes of the Hexoskin device relative to other devices in certain instances may be attributable to potential corruption or variations in the data collected from the wearable device, or due to the random selection of imposter and genuine samples during the testing process of each case. When we look at the performance of this device in Phase 1, the EER values vary between 26.04% and 18.27% in the NB classifier and between 31.67% and 25.42% in the DT classifier. Although it was observed that the results decreased significantly for both classifiers in Phase 3, the NB classifier showed much more successful results.

6.1.2 Emotional Status-aware Models

The RespiBan device is a consumer-based wearable device. In addition to our comparative studies between wearable and medical devices, it is crucial to examine the impact of various emotional states on biometric verification, in order to clarify the applicability of our models to diverse datasets. Architecture (shown in Fig.6.1), all models and parameters used in the emotional status-aware biometric verification are the same as those applied in activity-aware models (i.e. Section 6.1.1) for the performance comparison.

Table 6.5 shows the biometric verification performances in terms of EERs for both Manhattan and Euclidean distances-based features in each emotional status and classifier case. The best EER results are indicated in bold.

TABLE 6.5: Biometric verification performances of RespiBAN device in terms of EER. 50% of genuine samples were used in testing.

			DT		NB	
	Feature Names		**Manhattan**	**Euclidean**	**Manhattan**	**Euclidean**
Activity Names	**Baseline**	**Tr-80%**	15.08%	15.12%	**8.32%**	9.43%
		Tr-60%	15.01%	15.05%	**8.32%**	9.44%
		Tr-50%	**14.88%**	15.18%	8.48%	9.30%
	Stress	**Tr-80%**	**15.01%**	15.43%	8.41%	9.40%
		Tr-60%	15.06%	15.22%	**8.33%**	9.49%
		Tr-50%	15.20%	15.12%	8.34%	9.33%
	Amusement	**Tr-80%**	15.03%	15.18%	8.42%	9.43%
		Tr-60%	14.99%	15.02%	**8.33%**	9.51%
		Tr-50%	**14.67%**	15.04%	8.40%	9.35%

Based on the results, feature sets using Manhattan distances had the lowest EERs. In addition, NB achieved better performances than the DT classifier in all cases. In Phase 2 (i.e. Chapter 5) emotional status classifications, it was observed that the accuracy of classification decreased as the training set size was reduced from 80% to 50%. However, in the biometric verification model based on this classification, the *Tr-50%* and *Tr-60%* cases typically yielded the lowest EER results (i.e. the most successful results). This situation can be explained by the increased number of samples utilised in biometric verification, as the total number of samples passing

from the activity classification stage to biometric verification increases. The use of a larger quantity of samples in the biometric verification training process often results in lower EER values.

Although the minimum EER value for the NB classifier is 8.32%, which was achieved for the baseline emotional state, it was nearly equivalent to the minimum EER values obtained for the stress and amusement cases using this classifier. However, the minimum EER value achieved using the DT classifier was 14.67%, obtained for the amusement emotional state. The optimal values for both classifiers were achieved using features based on Manhattan distances.

Although this device obtained higher EER results in Phase 1 and lower activity classification accuracy percentages in Phase 2 than the devices in the Vollmer dataset, it proved that emotional status classification had an effect on biometric verification in Phase 3. The EER results in Phase 1 ranged from 30.67% to 19.81% for the NB classifier, and between 36.43% and 26.33% for the DT classifier. In Phase 3, the EER results were observed as approximately 15% for DT and approximately 9% for NB.

Considering the proposed framework, the optimum conditions for an emotional status-aware biometric verification model are classifying the samples in the *Tr-60%* case using Manhattan distance features and the NB classifier. Although the baseline emotional status achieved a worse classification rate than other emotional states in Phase 2, EER performances under the optimum conditions specified in Phase 3 were almost equal for all emotional statuses. This shows that the proposed framework achieves more successful results for the baseline emotional status.

6.2 Deep Learning Models for Activity-aware Biometric Verification

As a general framework, as in ML (i.e. Section 6.1), we aimed to first classify activities and then create biometric verification models for each activity in DL models.

To employ the same approach, we classified activities using GoogleNet, ResNet50, and DenseNet201 CNN structures and then create biometric verification models for each activity using the same CNN structures. However, the accuracy rates achieved in Phase 2 through ML were not replicated in DL models. There could be several factors contributing to this, such as incompatibility with the desired activity classification task with pre-trained CNN models for the classification of objects and the images selected not being distinct enough for the activities.

Utilising biometric verification with the newly assigned class label, as we do in ML, regardless of whether the samples are classified incorrectly or correctly, results in more errors and an unrealistic scenario in DL models. This is due to the insufficient accuracy rates obtained from DL models in Phase 2 and the negative impact of excessive errors in activity classes on our ability to analyse the influence of each activity on biometric verification. Therefore, scalogram, spectrogram and Mel-spectrogram images were manually divided into real activity classes and biometric verification was made with the ResNet50 and DenseNet201 CNN models, which provided the best results in Phase 2. In this way, it was assumed that a hypothetical activity classification model with a 100% accuracy rate moves to Phase 3.

Since this study is a follow-up phase of Phase 1, all hyperparameters and procedures were identical. To summarize these hyperparameters, 11 of 13 subjects are reserved for training and 2 for testing. For both CNN structures, the mini-batch size is set to 32, the maximum number of epochs is set to 4, the initial learning rate is 0.0003, and the execution environment is set to CPU. During Phase 3, we chose not to use pre-trained models because they didn't yield satisfactory results for our task. Instead, we continued with the ResNet50 and DenseNet201 models (see Sections 4.2.1 and 4.2.2) that we had trained from scratch during Phase 1. Table 6.6 includes information on the number of images used for training, validation, and testing in 2, 4, and 10 seconds time windows for both CNN structures.

This study is important to see if using activity classification can improve the direct biometric verification model from Phase 1. Moreover, the study aims to

TABLE 6.6: The number of images in training, validation and testing sets for each activity

	2 sec	4 sec	10 sec
# of Images in Training	1254	748	330
# of Images in Validation	308	187	88
# of Images in Testing	170	102	46

investigate the impact of different activities on biometric verification. Therefore, these studies focused on examining the performance of Faros and Hexoskin devices, which demonstrated the best results during ML models.

In this section, we assess the biometric verification performance of data categorized according to different physical activities. We utilise ResNet50 and DenseNet201 CNNs for each activity and each device. Subsection 6.2.1 details the experiments conducted using the ResNet50 CNN model, while Subsection 6.2.2 explains the experiments conducted using the DenseNet201 CNN model.

6.2.1 ResNet50 CNN

In line with Phase 1, the DL parameters defined in Section 6.2 were employed during Phase 3. Furthermore, the ResNet50 model, as described in Fig.4.4, was created for the purpose of direct biometric verification and it was utilised separately for each activity.

Table 6.6 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.

Similar to Phase 1, we selected pairs *P1-P2, P6-P7, P8-P9* and *P11-P12* to be tested individually as unseen subjects while the other 11 subjects were used for

the system training. Once the training process is finished, the model that has been trained (i.e. *DAGNetwork*) is saved. As illustrated in Fig.4.4, the final three layers are removed to enable biometric verification. These particular layers were designed for training purposes and are not intended for biometric verification. The new model, which is created to be output from the normalisation layer, is saved for testing as another *DAGNetwork*. Using the MatLab "*predict*" toolbox, an embedded model is created within this *DAGNetwork*. This embedded model is tested by being given a verification task to evaluate its performance on unseen subjects. It uses either 1, 3, or 5 (i.e. N numbers) images for each subject to create a user template for enrollment. The template is generated by averaging the embeddings obtained from N genuine samples of the same subject. The mathematical formula of the template creation is expressed in Eq.(4.1).

To create templates, a random number of N samples is selected from each person's embeddings. Following this, the model is tested using both genuine and imposter samples from the same and different subject classes. The verification response is determined by calculating the Euclidean distances between the verification embeddings and user templates. A positive response is given if the distance is below a set threshold, and a negative response is given if it is above the threshold.

6.2.1.1 Faros Device

The ResNet50 CNN model, whose hyperparameters, training and testing procedures were explained in the previous sections, was applied to the ECG signals obtained from the Faros device for biometric verification. Each activity was evaluated separately with different enrollment times and several image types in activity-aware biometric verification tasks. Mean EERs are shown for different enrollment times and three image representations in Table 6.7. The best performances of each activity and each image type are presented in bold.

The results can be examined separately for each time-frequency representation and each activity. When analysing the effectiveness of time-frequency representations, it was found that the scalogram had the lowest EER results (i.e. it achieved the

TABLE 6.7: Biometric verification performances of Faros device in terms of EER for each activity.

Activities	Resting			Walking			Standing			Uphill Walking		
Image Types \ # of Genuine Samples	**1-S**	**3-S**	**5-S**	**1-S**	**3-S**	**5-S**	**1-S**	**3-S**	**5-S**	**1-S**	**3-S**	**5-S**
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%**

best performances), followed by the spectrogram in the second place, and the Mel-spectrogram in last place. The optimal results for scalogram images are achieved with 4 second time windows, followed by 10 second time windows, and then 2 second time windows. Furthermore, cases containing *3-S* and *5-S* genuine samples generally outperformed *1-S* cases while containing similar EER results. Although the lowest EERs are obtained from the *walking* activity, when we examine all cases in general for scalogram and Mel-spectrogram images, the best performances can be listed as *resting, standing, walking* and *uphill walking*, respectively.

The best results in spectrogram images are achieved using time windows of 10 seconds, followed by 2 seconds, and then 4 seconds. Additionally, cases with *5-S* genuine samples tend to have better EERs than those with *3-S*, while cases with *1-S* samples generally have the highest EER results. The *walking* activity during *SP4* and *SP10* cases obtains the lowest EERs, but when looking at *SP2* cases, the top performances are seen in *resting, standing, walking* and *uphill walking*, respectively.

The results indicate that the optimal performance of Mel-spectrogram images is achieved through the utilisation of time windows with durations of 10 seconds, 4 seconds, and 2 seconds. Furthermore, it was observed that cases employing *3-S* and *5-S* genuine samples usually had lower EERs than those utilising *1-S* samples

while having similar EER outcomes. Similar to scalogram images, in terms of physical activity, the lowest EERs were achieved during *walking*.

During this study, the impact of short enrollment times was explored. While it is commonly known that EER decreases with long enrollment times, it was found that also with *3-S* cases, successful results were achieved in this study. If we compare the obtained results with the EERs in Phase 1, the mean EER of all activities in Phase 3 is lower than in Phase 1. However, if analysed on an activity basis, EERs for resting, walking and standing activities in all cases are lower than Phase 1. Some cases with uphill walking activity achieved higher EER results from Phase 1. The overall trend is that EER results for both phases 1 and 3 decrease from 2 second cases to 10 seconds and from *1-S* cases to *5-S*.

6.2.1.2 Hexoskin Device

The ECG signals obtained from the Hexoskin device were subjected to biometric verification using the ResNet50 CNN model, whose hyperparameters, training, and testing procedures were thoroughly explained in sections 6.2 and 6.2.1. The activity-aware biometric verification tasks were evaluated separately, considering different enrollment times and various image types. Mean EERs are shown for different numbers of genuine samples and three image representations in Table 6.8 for each activity case.

Upon examination of the results, it was determined that the effect of each time-frequency representation and activity could be evaluated separately. In terms of time-frequency representation effectiveness, the scalogram was found to have achieved the most successful outcomes, as indicated by its lowest EER results. As in the Faros device, the spectrogram images provide the second most accurate results, while the Mel-spectrogram was found to have the least successful performance.

While the most successful results were obtained in the 4 second time windows in the scalogram images, there are 10 second time windows in the second place and

TABLE 6.8: Biometric verification performances of Hexoskin device in terms of EER for each activity.

Activities	**Resting**			**Walking**			**Standing**			**Uphill Walking**		
Image Types \ # of Genuine Samples	**1-S**	**3-S**	**5-S**	**1-S**	**3-S**	**5-S**	**1-S**	**3-S**	**5-S**	**1-S**	**3-S**	**5-S**
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%

2 second time windows in the poorest performances This situation is different for spectrogram and Mel-spectrogram images. The 10 second time windows for these images yield the lowest EERs, followed by the 4 second time windows and then the 2 second time windows cases.

For almost all cases, an increase in EERs is observed as the enrollment time is shortened. The *standing* activity showed the lowest EERs across all time-frequency representations. If we look at the performance of activities in general for all image types, EERs increase in standing, resting, walking and uphill walking activities, respectively (from the lowest to the highest one). The minimum EER was obtained as 0.49% from the *SC4* standing case. This value is the mean value of the selected 4 subject pairs (i.e. *P1-P2, P6-P7, P8-P9* and *P11-P12*), but if they are examined one by one, it is obtained as FP=1, TN=101, FN=0 and TP=102 for each subject pair. In this way, FAR=0.49% and FRR=0% were calculated.

When we compare the results obtained from Phase 3 with the EERs from Phase 1, we can see that the EERs for all activities in Phase 3 are generally lower than they were in Phase 1. The number of training and validation images in Phase 1 is higher than in Phase 3. Even in that case, the proposed framework shows better performance than direct biometric verification. Although Hexoskin and Faros devices achieved very close results in Phase 3, as in Phase 1 and Phase 2, it

was observed that Faros device was slightly more successful in general performance than Hexoskin device in ResNet50 model.

6.2.2 DenseNet201 CNN

During Phase 3, the DL parameters outlined in Section 6.2 were implemented in accordance with Phase 1. Additionally, a DenseNet201 model was developed, as indicated in Fig.4.5, specifically for direct biometric verification. This model was used separately for each activity.

The number of images specified in Table 6.6 was used for training, validation and testing. The subject pairs specified in Phase 1 and subsection 6.2.1 were also used for testing purposes as unseen subjects in the DenseNet201 model and the system training utilised the other 11 subjects. The general biometric verification model in Phase 3 using the DenseNet201 CNN structure is the same as stated in Subsection 4.2.2.

As in the ResNet50 structure, Faros and Hexoskin devices were examined separately in subsections 6.2.2.1 and 6.2.2.2, respectively in the DenseNet201 structure.

6.2.2.1 Faros Device

We utilised the DenseNet201 CNN model on the Faros device to assess the performance of activity-based biometric verification. To evaluate each activity, we used different enrollment times and various image types. Table 6.9 displays the mean EERs for the different enrollment times and three image representations. In the Table, the states of *1-S, 3-S*, and *5-S* indicate varying enrollment times. The *1-S* case denotes the shortest enrollment time, while the *5-S* case represents the longest enrollment time used in this study.

In this model, the lowest EER results were obtained from the *walking* activity for all image types, unlike other models. However, in general, it was observed that

TABLE 6.9: Biometric verification performances of Faros device in terms of EER for each activity.

Activities	Resting			Walking			Standing			Uphill Walking		
Image Types \ # of Genuine Samples	1-S	3-S	5-S	1-S	3-S	5-S	1-S	3-S	5-S	1-S	3-S	5-S
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%

from the lowest mean EER of each image type to the highest, *resting, walking, standing* and *uphill walking* activities were achieved, respectively.

As a common trend with other models, the most successful results were obtained from scalogram, spectrogram and Mel-spectrogram images, respectively. However, if we examine each time-frequency representation separately, it is observed that the time windows show different trends than the other models. As an example, the order of EER from lowest to highest in scalogram images is 4 seconds, 10 seconds, and 2 seconds. However, in spectrogram images, it is 10 seconds, 4 seconds, and 2 seconds. Additionally, Mel-spectrogram images have an order of 10 seconds, 2 seconds, and 4 seconds. Therefore, in general, we can say that the longest enrollment time usually achieves lower EER results. However, another factor affecting the enrollment time is the number of genuine samples used. It is observed that *1-S* and *5-S* cases are quite successful for all image types. 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-S* case since its discrimination might be higher among other images. These results are important to see the effects of enrollment times and time windows on different time-frequency representations and several activities.

After comparing the results, it can be concluded that the ResNet50 model is

slightly more successful than the DenseNet201 model for the Faros device. Although two different CNN models obtained the lowest EER results from *walking* activity on the same device, they were successful at different time windows and different genuine sample cases.

If we compare the Faros device with the DenseNet201 CNN results in Phase 1, Phase 3 achieved lower EER results in all cases. While some results in the Phase 3 ResNet50 model give higher EER results than Phase 1, the absence of this situation in the DenseNet201 model indicates that the DenseNet201 structure is better suited to this device.

6.2.2.2 Hexoskin Device

We conducted an assessment of activity-based biometric verification using the Hexoskin device. Different genuine samples and three image types were used to evaluate each activity. The mean EERs for different enrollment times and different image representations are shown in Table 6.10. The Table includes the enrollment times of *1-S, 3-S*, and *5-S*. The lowest EERs for each activity and each image type are indicated in bold.

TABLE 6.10: Biometric verification performances of Hexoskin device in terms of EER for each activity.

Activities	**Resting**			**Walking**			**Standing**			**Uphill Walking**		
Image Types \ # of Genuine Samples	**1-S**	**3-S**	**5-S**	**1-S**	**3-S**	**5-S**	**1-S**	**3-S**	**5-S**	**1-S**	**3-S**	**5-S**
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%**

After examining the Table, it appears that the results are consistent with other devices and the Resnet50 model overall. For instance, scalogram images have proven to be the most effective, followed by spectrogram images, and Mel-spectrogram images have had the least success (i.e. highest EERs). Although *Sc4* images have achieved the lowest EER results, it has been noticed that the EERs tend to increase as the time window size (in other words, from 10 seconds time windows to 2 seconds time windows) decreases in general structure. Moreover, the *standing* activity yielded the lowest EER results across all time-frequency representations.

Upon analysis of the scalogram images, we see that contrary to the general trend, the lowest EERs are obtained from the 4 seconds time windows, followed by the 10 seconds and 2 seconds time windows. In addition, if we look at the *Sc4* case with the best performances, it seems that contrary to the general trend, the best results are obtained from the *1-S* cases, which represent the shortest enrollment time. It is followed by *3-S* and *5-S*, respectively. However, when looking at the scalogram images in general, this situation differs from best to worst as *5-S*, *3-S* and *1-S*, respectively. If we examine the lowest EERs by activity, *resting* and *standing* activities show the best results for the lowest enrollment time, while for longer enrollment times, this is *standing, walking, resting*, and *uphill walking*, from best to worst. The best performance of 0.49% EER is the same for all enrollment time cases. In addition, 0.49% FAR and 0% FRR were obtained in all cases.

When analysing spectrogram images, it has been noticed that longer enrollment times lead to reduced EER outcomes. Moreover, when it comes to the overall performance of activities, *walking, standing, resting*, and *uphill walking* are ranked in descending order of success. When the Mel-spectrogram images are taken into account, it was observed that *3-S* cases also showed successful results like *5-S* cases. In addition, while it was observed that the lowest EERs for the *Mel2* case were obtained from *resting* activity, this situation was observed as *walking, resting, standing* and *uphill walking*, respectively, from the lowest EER to the highest.

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 for the Hexoskin device. From this observation, it can be concluded that the DenseNet201 model is more effective when used the wearable device's data.

If we compare the results obtained from the Hexoskin device and DenseNet201 CNN in Phase 1, we can see that Phase 3 has lower EER results in almost all cases. In the case of uphill walking, the *Sp10* case achieved higher EERs in Phase 3 than in Phase 1. However, if we look at the mean performances of all activities in Phase 3, we can observe that Phase 3 outperformed Phase 1.

6.3 General Inferences from Deep Learning Models

After examining ResNet50 and DenseNet201 CNN structures, which are known to be highly effective DL models, we tested various parameters. Phase 3 consistently achieved more successful results than Phase 1 in both cases. Table 6.11 displays the mean EERs we acquired from the DL model during Phase 3. Comparing Phase 1 and Phase 3 is simpler when using the mean values of all activities. Additionally, it enables us to easily interpret the improvement in the performance of both medical and wearable devices.

In all instances, the mean EER values in Phase 3 are lower than those obtained in Phase 1. The difference between Phase 3 and Phase 1 is most significant in spectrogram and scalogram images when using minimum sample cases (which are 2 seconds time windows and 1-S 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 Phase 3 and Phase 1 EERs is highest when using 5-S genuine sample cases and 10 seconds 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

TABLE 6.11: The mean EER performances of DL models in Phase 3

	ResNet50						DenseNet201					
	Faros			Hexoskin			Faros			Hexoskin		
Image types	1-S	3-S	5-S	1-S	3-S	5-S	1-S	3-S	5-S	1-S	3-S	5-S
Sc2	24.96%	24.49%	25.37%	25.93%	24.31%	22.82%	24.71%	25.77%	25.19%	22.83%	20.85%	19.52%
Sc4	8.0%	7.08%	6.59%	3.53%	3.09%	3.12%	7.26%	7.63%	6.62%	2.82%	3.07%	3.19%
Sc10	15.88%	14.01%	13.57%	17.16%	14.44%	13.90%	15.79%	14.60%	13.51%	13.05%	11.82%	11.01%
Sp2	30.19%	28.64%	28.46%	33.92%	31.49%	31.01%	31.76%	31.40%	30.0%	29.41%	28.39%	27.87%
Sp4	30.69%	28.62%	28.39%	23.75%	22.06%	20.74%	26.99%	27.04%	26.72%	21.93%	21.52%	20.27%
Sp10	19.42%	16.75%	16.07%	16.13%	13.61%	12.02%	20.23%	18.33%	17.25%	11.81%	10.58%	10.04%
Mel2	33.02%	34.30%	34.41%	38.18%	36.77%	36.29%	33.42%	34.03%	33.86%	37.33%	36.29%	36.88%
Mel4	31.50%	31.01%	32.48%	32.56%	32.76%	33.0%	32.97%	33.52%	33.51%	34.52%	33.35%	32.79%
Mel10	24.21%	22.01%	21.90%	25.16%	23.06%	22.67%	22.42%	22.56%	23.24%	22.53%	21.89%	20.47%

size, more information can be obtained from the image, resulting in better results in Phase 3 during analysis.

6.4 Discussion

In Phase 3, the proposed activity-aware biometric verification framework is described. To examine this framework, we utilised several ML and DL models on different mobile devices and different datasets. This chapter shows that ECG signals classified according to physical activities or emotional states have low error rates even at lower enrollment times when used for biometric verification.

When analysing ML models, we assess the impact of generated Manhattan distance-based and Euclidean distance-based features on biometric verification across four distinct devices. We conduct this examination at varying rates of testing samples that pass activity classification. For instance, during Phase 2, the samples were divided into 80% for training and 20% for testing (i.e. *Tr-80%* case). In Phase 3, we used the portion that was tested in Phase 2. In other words, as explained in Subsection 6.1.1.1, 237 samples were able to proceed to Phase 3 in the *Tr-80%*

case, while 568 samples were able to proceed to Phase 3 in the *Tr-50%* case for the Faros device. Among these conditions, it was observed that the *Tr-50%* cases have a higher number of samples in Phase 3. The utilisation of more samples in Phase 3 resulted in lower EER outcomes for both activity-aware and emotional state-aware biometric verification. When examining the mean EER values of both feature sets in Phase 3, in the activity-aware system, the *Tr-50%* case showed 0.61% more successful results in the NB classifier and 1.02% in the DT classifier than the *Tr-80%* case. In addition, in the emotional states-aware system, the *Tr-50%* case showed 0.04% more successful results in the NB classifier and 0.13% in the DT classifier than the *Tr-80%* case. It is noteworthy that these findings are significant for improving biometric verification accuracy because more samples mean more enrollment times.

It is not possible to make a direct comparison since the two datasets were collected under different conditions. However, it is generally observed that biometric verification results are more successful in samples that contain physical activity compared to those that contain emotional status. When examining the devices in the Vollmer dataset that were collected under identical conditions, the Faros device ranked as the most successful, followed by Nexus, SomnoTouch, and Hexoskin in descending order of success. According to the results in Phase 3, the performance of medically approved devices is better than that of the wearable device. However, when we consider the device performances in Phase 1 and Phase 2, it is seen that the Hexoskin device is as successful as the Faros device.

Examining the performances of physical activities and emotional states one by one is important in terms of testing the applicability of the features and models used in real life. The physical activity that resulted in the lowest EER was *resting*, while *amusement* and *baseline* emotional states had similar EER results. Furthermore, while the utilised features have been observed effective in the emotion-aware biometric verification model, there are noticeable similarities in the results when analysing individual emotional states. This shows that physical activities cause more changes in ECG signals and classifying physical activities contributes more to biometric verification. In this study, activities involving less movement (i.e.

resting and standing) were more successful in both Phase 2 and Phase 3 than activities involving more movement (i.e. walking and uphill walking). Studies using other ML models also support this observation [70, 92]. Although some of the ML methods tested in the study [70] gave better results than our study in different body postures, they obtained worse results than our study in both activity classification and biometric verification in the case involving exercise.

In ML models, the proposed framework in Phase 3 has improved the EERs in Phase 1. Upon analysing the overall performances, it appears that the EERs are higher than those reported in some studies in the literature [61, 93, 99]. One possible explanation for this could be the limited number of features that were utilised. However, this study focused on only Manhattan and Euclidean distances-based features to compare their performances. The significant difference between Phase 1 and Phase 3 EERs, and the very high classification accuracy rates in the activity classification in Phase 2 show the success of the proposed framework and ML models used. At the same time, exploring the effects of different classifiers using the most known features on the E-HOL and WeSAD datasets in Phase 1 guided the following phases to determine the correct classifiers and different enrollment times effects.

The proposed framework was evaluated utilising DL models, specifically ResNet50 and DenseNet201 CNNs. However, due to the COVID-19 pandemic, only the best performing medically approved Faros device and the wearable Hexoskin device were compared as in Phase 1. Despite using fewer training images in Phase 3 compared to Phase 1, there was a notable reduction in EER outcomes.

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 [154], CNN+LSTM [255] and ECGXtractor [11] CNN models. Labati et al. [154] used 3 leads ECG data from E-HOL dataset to train the CNN model. They used 300 seconds data per person to train the CNN model and achieved

3.37% EER from lead X, 4.86% EER from lead Y and 4.15 % EER from lead Z using 8 heartbeats for enrollment. Martin et al. [255] used different training,validation and enrollment time (from 15 seconds or 75 samples to 45 seconds or 225 samples) cases to see the effects of them on their novel BioECG CNN model and they achieved from 0% to 10.13% EERs for different cases. Melzi et al. [11] created new ECGXtractor CNN model and they trained it with 12 leads ECG data from In-house dataset. They achieved EERs across various datasets, with values spanning from 0% to 7.97% using 3 single segments (i.e. 3 hearbeats) of enrollment data. Although these studies give similar results to our study, when their enrollment times are taken into consideration, it is determined that they use more enrollment time than our study (we used from 2 to 50 seconds enrollment time).

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 categorize 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 etc.), it is important to investigate short enrollment times so that this process can be done quickly and with less error.

Similar to our studies, Nawawi et al. [100] used the quadratic-SVM classifier that performs biometric verification separately for walking, standing and sitting activities using the data obtained from the Hexoskin device. In the study [100], the number of genuine and imposter samples was not specified, but it was stated that 80% of data for training and 20% of data for testing were used. The approximate EER values calculated from the published FAR and FRR ratios are higher than our study. Moreover, there is a significant difference between the FAR and FRR ratios, resulting in a low level of reliability for the biometric system.

Byeon et al. [83] conducted a comparative analysis of various CNN architectures

and time-frequency representations for biometric identification. Their findings revealed that the performance of the CNN models varied 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. 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 [83] supports our study in these points, it differs from our studies in that it does not include activities and does not perform biometric verification.

While the medically approved Faros device gave better results in the ResNet50 model in general, the Hexoskin device was found to be more successful in the DenseNet201 model. The activity-aware biometric verification study has proven that it is a very reliable model for both wearable devices and medical devices, with **0.49%** FAR, **0%** FRR and **0.49%** EER outcomes.

Although the best results of biometric verification tasks were achieved from DL models, if we compare the performances of ML and DL models in this thesis, scalogram images in Phase 1 and all time-frequency representations in Phase 3 (especially cases with 10-second windows) gave better results than the ML model in some cases. When the enrollment time was increased in the DL model, the obtained EER results were generally lower compared to the EER results in the ML model. However, in Phase 2, ML outperformed DL for all cases. The main reasons for this situation are the shorter enrollment times used in the DL model, while the longer enrollment times in the ML model for the investigation of feature performances and while the DL model should be trained with a much larger number of images, it should be trained with fewer images.

Chapter 7

Conclusion and Future Work

7.1 Summary

ECG signals have been frequently studied in health monitoring and authentication applications as a biometric trait, as mentioned in Chapter 2. Although ECG biometrics have been researched in different fields, it is still not as widely used as fingerprint and face recognition. A few of the main reasons for this are that ECG recorders/sensors are less common than fingerprints and the need for more templates and data for the biometric verification system. ECG signals are affected by factors such as emotional status, physical activities and health conditions. Even if the general structure of the heartbeat does not change by these factors, the duration of heartbeats and the amplitude values of Q, R and S peaks can change. Each change affects the biometric verification model. To reduce these effects on verification models, more templates are needed. More templates need more storage area and are more time-consuming during verification.

In this study, the performances of mobile devices are compared with ML and DL models in order to investigate the reasons mentioned, and a new framework is presented. In our context, portable ECG recorders are described as mobile devices. This study investigated the quality of the collected signal and device by comparing the performance of different devices and concluded that wearable

devices can perform as well as medical devices. It also investigated the effects of emotional states and physical activities on biometric verification models by classifying different activities/emotions, and it was concluded that both factors increase the error rate of the model, and the new framework was introduced by creating a reliable model using less amount of templates and data by using short enrollment times for different models.

In the upcoming sections, we will mention our research questions and the responses we have obtained through our studies. Additionally, we will consider areas that require further investigation due to certain limitations.

7.2 Research Findings

In this study, experiments were conducted to understand and solve some of the problems preventing the more widespread use of ECG biometrics on consumer-based ECG recorders.

– **What is the baseline equal error rate from various devices using ECG?**

To determine the baseline equal error rate of ECG signals from different devices, we began our study by examining large datasets for real-life accuracy and diverse device types. Data collection procedures were different as we used open-source datasets. Since this situation creates different noises in ECG signals, different methods have been applied in the signal pre-processing section. Although a uniform method could not be applied, this gave us the opportunity to research several devices in different experimental environments. In addition, applying the same model to different datasets and comparing the results increased the reliability of the study.

According to the results obtained from our study explained in Section 4.1.1, the effect of enrollment times on different classifiers was investigated. It concluded that

NB and DT outperformed the other classifiers. Due to its larger amount of subject and longer recording duration (i.e. 24 hours), the E-HOL dataset achieved higher EER results compared to the WeSAD dataset. Even in this case, it achieved close EERs on both the medical device and the wearable device we tested. This study tested the applicability of the same biometric verification model and parameters to both wearable and medical devices and showed that longer enrollment time resulted in lower equal error rates.

As we examine other research that utilises the E-HOL dataset, it has been demonstrated that the results vary depending on whether the samples tested in the biometric verification model are selected from a long or short time period [154, 166]. In study [154], the EERs obtained using a single-channel ECG signal are similar to our results in the case where they chose a short time period (300s). In the case they chose a long time period (150min and 500min), they achieved higher EERs than our results. In addition, the EERs (EER: 7.01%, 8.58% and 12.98%) obtained in the study [256] conducted with samples of 140 participants selected from the 300s period were higher than those obtained in the 50s enrollment time case in our study. In our study, the baseline verification rates have ranged from 3.64% to 6.30% for the medical device, and between 3.02%-4.57% for the wearable device.

In our study, where the new features explained in Section 4.1.2 were introduced, the baseline verification performances of 5 devices were compared using NB and DT classifiers. This study supported the previous study in terms of an increase in the number of enrollment samples resulting in a reduction in the baseline of the EERs. In addition, as a general trend, features containing Manhattan distances were more successful than those containing Euclidean distances in all cases. In the obtained results, the Faros device demonstrated a notable average EER of 21.72% in all cases among medical devices. Similarly, the Hexoskin device achieved an average EER of 22.08% among wearable devices. In addition, the fact that the dataset containing physical activity obtained more successful results than the dataset containing various emotional states formed the basis for our DL studies.

If we compare the WeSAD dataset, which is common to both studies in Sections 4.1.1 and 4.1.2, for two separate feature sets, higher EER results were obtained for the feature sets in Section 4.1.2 than the features described in Section 4.1.1, despite the increase in the enrollment time used. The reasons for this are that the number of features decreased from 15 to 3 (3 for the Manhattan distance, 3 for the Euclidean distance), the time windows decreased from 10 seconds to 4 seconds and different filtering methods in pre-processing. This study showed that using a shorter time window can also be useful in classical ML. In addition, we have demonstrated that wearable devices that achieve very similar results in terms of biometric verification performances can be used instead of medical ECG recording devices that are impractical to use in real life.

Considering the DL model, it shows the same trends as the other two studies in terms of increasing enrollment time and decreasing EER results. In terms of the best performances, the investigation utilising DL models yielded higher EERs compared to the study in 4.1.1, yet lower than the study in 4.1.2. However, the DL model was tested at enrollment times shorter than 50 seconds, which the ML model cannot achieve. In addition, it is not possible to make a direct comparison between the results of the E-HOL and WeSAD datasets collected under different conditions and the Vollmer dataset. Therefore, if we compare the ML and DL models using the same Vollmer dataset, it is seen that using the DL model improves the results. Although the DL model needed more training data for learning and creating its features, it was also able to achieve low EER results for much shorter enrollment times than ML. Thus, it is a more suitable candidate for a real-life biometric verification scenario.

– **Can we accurately detect physical activity or emotional status from ECG waveforms?**

As seen in the literature, many methods have been used for activity classification. However, we have addressed this question differently in ML and DL models. Our approach in ML was to compare the performance of the features we produced

on different classifiers and several devices. Moreover, since our study allowed the comparison of different devices and datasets using the same evaluation method, it was an approach that contributed a solution one of the open challenges in the literature [257]. According to the results obtained, classifiers such as KNN and DT, known as lazy/weak learners, achieved an accuracy rate of approximately 65% in physical activities and 60% in emotional states, while ensemble classifiers achieved an average of 95% in physical activities and 85% in emotional states. In light of our proposed features achieving high accuracy rates, we affirmatively addressed our research question.

Our study analysed the produced features among themselves in terms of classification performance and found that those based on the Manhattan distance were more successful than those based on the Euclidean distance. Additionally, we observed that activities with more body movements had a lower accuracy rate due to signal deterioration. When it comes to emotional states, we found that stress and amusement are easier to classify than the neutral/baseline state. Based on these results, it can be concluded that emotional states in which the heart rhythm changes but the obtained signal quality does not change can be better classified than the baseline emotional state. If we examine the overall accuracy rates, we observe that the dataset containing physical activities is classified better than the dataset containing emotional states. Possible reasons for this situation are that the physiological responses of physical activities are predictable (e.g. exercise causes increased heart rate, which can be detected in ECG patterns), consistent (activities often involve repetitive movements, e.g. pedaling, walking) and can be measured objectively more easily. Considering emotional states, unlike physical activities, emotions lack specific ECG patterns. Emotional states exhibit considerable variability across individuals and are subject to contextual, personality, and cultural influences. In addition, emotions consist of complex neural and autonomic responses. ECG signals may exhibit non-linear features related to emotional changes, making classification challenging [251].

When we examined the issue in the context of the DL model, to build a reliable model with high accuracy rates, we compared several DL parameters. The general

trend is more samples and more epochs achieved higher classification results. No significant difference was observed between the performances of optimisers, but the Adam optimiser showed slightly better results. The results of the DL model reached a maximum classification accuracy of 70%. This rate is below the classification rate achieved in ML. This observation suggests that improving the transfer learning technique within the deep learning model requires adjusting parameters or training the CNN model with task-specific data.

– **Do physical activity or emotional status classification prior to biometric verification improve performance in medical and wearable ECG devices?**

This question involves investigating an open challenge that also affects the extensive use of ECG biometrics in real life. ECG is a signal that becomes unstable in the long term, depending on emotional status and physical activities [257]. A reliable biometric verification framework is proposed in which we can use the same evaluation method for different types of devices and is not affected by the stability of the signal.

When we evaluated the results in Phase 1 and Phase 3 together, we noticed that the proposed biometric verification framework achieved significant performance improvements on all types of mobile devices. One of the primary reasons for this is that during Phase 1, the genuine and imposter samples were chosen randomly, while in Phase 3, each sample was selected based on its specific activity or emotional state. As a result, EERs are higher during Phase 1 because samples can be selected randomly from different physical activities or emotional states. For example, while the samples selected to create the template were selected from the resting activity, the samples selected for testing may have been selected from the running activity. This may have caused the template and test data to match each other with more errors.

Upon analysing the emotional status case, it was noted that the Phase 3 results, which involved the use of the DT classifier, were around half of the EERs obtained

in Phase 1. When analysing the situation using the NB classifier, the results of Phase 3 are approximately one-third of the results obtained in Phase 1. Upon analysing the physical activity case, the same pattern was observed in the case that included physical activity classification prior to biometric verification. It is known that NB performs better than DT in high-dimensional data. However, even with fewer samples in Phase 3, DT still performed worse than NB. This indicated that the utilised features were more distinguishable when employing the NB classifier.

Biometric verification results were analysed for each device separately and their performances were compared in Chapters 4 and 6. Considering the overall mean results of the devices, classifiers showed more successful results in activities involving less body movement in Phase 3, while they achieved higher EER results in activities involving more body movement for the activity-aware system. In addition, in the emotional status-aware system, EER results were obtained lower in the baseline status than in stress and amusement. The situation is due to certain activities being easier to classify in Phase 2, such as resting and standing, compared to activities like walking and uphill walking.

Although the training data size decreased considerably in the Phase 3 DL model, obtaining successful results in low enrollment times in spectrogram and scalogram images shows the positive contribution of the presented framework to performance. With the advancements in technology, it is now possible to classify continuous activity from ECG recordings. This is especially useful for frequently used devices like smartwatches. Moreover, we proved that classified data can be used for biometric verification purposes, making it a valuable tool for various applications.

– **Do different machine learning models and their parameters contribute to the improvement of device performances?**

This research question covers all experimental chapters and the parameters were compared in detail across all sections. When examined from a biometric verification perspective, the following performance improvements were observed.

a. As the enrollment time used increased, more successful results were obtained because the system allowed it to learn more about the data.

b. As the number of subjects increases, performances tend to decrease in Phase 1. Although there is a performance decrease, the results obtained from a large dataset such as the E-HOL dataset are at an acceptable level.

c. Selecting training and testing data from the same or similar physical activities/emotional states increases performance (i.e. Phase 3 has better results than Phase 1).

d. Wearable devices can achieve as adequate performance as medical devices.

e. Scalogram images gave the most successful results because they are more compatible with real-life signals than the spectrogram [207].

f. Various CNN models yield varying outcomes based on the device type. For instance, in our case, ResNet50 achieved lower EERs on medical devices, whereas DenseNet201 obtained less EERs on wearable devices.

g. Although ML gives successful results, DL is a more usable model in real-life applications because it can be used in much shorter enrollment times.

When examined from the perspective of activity classification, linear and non-linear classifiers were tested. Non-linear classifiers and ensemble classifiers (consisting of many non-linear classifiers) have achieved more successful results than linear classifiers. The reason is that as the number of classes increases, linear classifiers have more difficulty in finding a hyperplane separating the classes. Although the use of more training data facilitated the more accurate classification of activity and emotional status during this stage, it resulted in diminished device performance in Phase 3 due to the increased demand for data.

In addition to examining the performances of each device, the optimum parameters were also examined for each activity and emotional status. In the proposed framework, the Manhattan feature set and the NB classifier are selected as optimum parameters for each activity and emotional status. In addition, since it was

observed that activities that involved less movement and neutral mood improved more in Phase 3, *standing*, *resting* activities and *baseline* emotional status were the most suitable characteristics for our proposed framework.

Since the number of training data was insufficient, a high classification accuracy could not be achieved in the DL model and much more time was required during the training period. For this reason, it has been concluded that ML models are more useful in activity classification for mobile applications.

7.3 Future Work

This study conducted extensive research on ECG biometrics across various mobile devices. The study addressed the research questions but also highlighted areas for further evaluation that could not be addressed due to constraints (see Section 1.4).

Due to the lack of ECG datasets containing physical activity or emotional states and the difficulty in collecting ECG data from different devices simultaneously, the proposed framework can continue to be examined extensively with other datasets. Examining the framework across a variety of devices and a wider range of activities or emotional states is essential for performance evaluation and advancement of wearable technology.

It is known that DL algorithms require more data, time and computation during the training process compared to traditional ML. The time-consuming training process is still a challenge for 2D DL algorithms. In this study, although many ML models and parameters have been examined and a biometric verification framework has been created according to these parameters, it is not possible to examine all parameters and all conditions. Therefore, other DL models can be compared within the framework presented.

7.4 Final Considerations

To conclude, we investigated ECG biometric verification models in this thesis and achieved multiple goals. The main notable contributions are listed below:

- In the academic literature, two novel features have been introduced, derived from the Manhattan and Euclidean distances of the Q, R, and S peaks. The performance of these features in biometric verification models and the classification of activity and emotional status has been extensively investigated.

- A novel activity-aware ECG biometric verification framework for mobile devices was introduced considering different activities, emotional status, and short enrollment times.

- By evaluating many types of devices under the same conditions, machine learning parameters were investigated in depth and the effects of these parameters on learning were observed.

The actual meaning of this study is to show that ECG biometrics can be widely used with wearable technology. In order to develop and improve wearable technology and device authentication security, an activity-aware system has been created and biometric verification errors that may arise from daily activities have been tried to be minimized in this study.

Bibliography

[1] Hazal Su Bıçakcı, Marco Santopietro, Matthew Boakes, and Richard Guest. Evaluation of electrocardiogram biometric verification models based on short enrollment time on medical and wearable recorders. In *2021 International Carnahan Conference on Security Technology (ICCST)*, pages 1–6, 2021. doi: 10.1109/ICCST49569.2021.9717372.

[2] Hazal Su Bıçakcı, Marco Santopietro, and Richard Guest. Activity-based electrocardiogram biometric verification using wearable devices. *IET Biometrics*, 12(1):38–51, 2023. doi: https://doi.org/10.1049/bme2.12105. URL https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/bme2.12105.

[3] Marcus Vollmer, Dominic Bläsing, and Lars Kaderali. Alignment of multi-sensored data: Adjustment of sampling frequencies and time shifts. In *2019 Computing in Cardiology (CinC)*, pages Page–1. IEEE, 2019.

[4] Lena Biel, Ola Pettersson, Lennart Philipson, and Peter Wide. Ecg analysis: a new approach in human identification. *IEEE transactions on instrumentation and measurement*, 50(3):808–812, 2001.

[5] Kiran Kumar Patro and P Rajesh Kumar. Effective feature extraction of ecg for biometric application. *Procedia computer science*, 115:296–306, 2017.

[6] ISO iso/iec 2382-37:2022 information technology — vocabulary — part 37: Biometrics. https://www.iso.org/standard/73514.html. Accessed: 2023-06-07.

[7] Anil K. Jain, Arun A. Ross, and Karthik Nandakumar. *Introduction to Biometrics*. SpringerLink (Online service), 2011.

[8] Anil Jain, Ruud Bolle, and Sharath Pankanti. *Biometrics: personal identification in networked society*, volume 479. Springer Science & Business Media, 1999.

[9] Anil K Jain, Arun Ross, and Salil Prabhakar. An introduction to biometric recognition. *IEEE Transactions on circuits and systems for video technology*, 14(1):4–20, 2004.

[10] Carlos Carreiras, André Lourenço, Hugo Silva, Ana Fred, and Rui Ferreira. Evaluating template uniqueness in ecg biometrics. In *Informatics in Control, Automation and Robotics: 11th International Conference, ICINCO 2014 Vienna, Austria, September 2-4, 2014 Revised Selected Papers*, pages 111–123. Springer, 2016.

[11] Pietro Melzi, Ruben Tolosana, and Ruben Vera-Rodriguez. Ecg biometric recognition: Review, system proposal, and benchmark evaluation. *IEEE Access*, 2023.

[12] Meryem Regouid, Mohamed Touahria, Mohamed Benouis, and Nicholas Costen. Multimodal biometric system for ecg, ear and iris recognition based on local descriptors. *Multimedia Tools and Applications*, 78:22509–22535, 2019.

[13] Ebrahim Al Alkeem, Chan Yeob Yeun, Jaewoong Yun, Paul D. Yoo, Myungsu Chae, Arafatur Rahman, and A. Taufiq Asyhari. Robust deep identification using ecg and multimodal biometrics for industrial internet of things. *Ad Hoc Networks*, 121:102581, 2021. ISSN 1570-8705. doi: https://doi.org/10.1016/j.adhoc.2021.102581. URL https://www.sciencedirect.com/science/article/pii/S1570870521001219.

[14] Tsu-Wang Shen, WJ Tompkins, and YH Hu. One-lead ecg for identity verification. In *Proceedings of the second joint 24th annual conference and*

the annual fall meeting of the biomedical engineering society][engineering in medicine and biology, volume 1, pages 62–63. IEEE, 2002.

[15] Steven A Israel, John M Irvine, Andrew Cheng, Mark D Wiederhold, and Brenda K Wiederhold. Ecg to identify individuals. *Pattern recognition*, 38 (1):133–142, 2005.

[16] Shihui Yin, Minkyu Kim, Deepak Kadetotad, Yang Liu, Chisung Bae, Sang Joon Kim, Yu Cao, and Jae-sun Seo. A 1.06-μw smart ecg processor in 65-nm cmos for real-time biometric authentication and personal cardiac monitoring. *IEEE Journal of Solid-State Circuits*, 54(8):2316–2326, 2019.

[17] Ettien Koffi. Voice biometrics fusion for enhanced security and speaker recognition: A comprehensive review. *Linguistic Portfolios*, 12(1):6, 2023.

[18] David Hernando, Nuria Garatachea, Rute Almeida, Jose A Casajus, and Raquel Bailón. Validation of heart rate monitor polar rs800 for heart rate variability analysis during exercise. *The Journal of Strength & Conditioning Research*, 32(3):716–725, 2018.

[19] Qardiocore. https://www.qardio.com/qardiocore-wearable-ecg-ekg-\monitor-iphone/. Accessed: 2023-08-15.

[20] Jesús Lázaro, Natasa Reljin, Md-Billal Hossain, Yeonsik Noh, Pablo Laguna, and Ki H Chon. Wearable armband device for daily life electrocardiogram monitoring. *IEEE Transactions on Biomedical Engineering*, 67(12):3464–3473, 2020.

[21] Erdong Chen, Jie Jiang, Rui Su, Meng Gao, Sainan Zhu, Jing Zhou, and Yong Huo. A new smart wristband equipped with an artificial intelligence algorithm to detect atrial fibrillation. *Heart rhythm*, 17(5):847–853, 2020.

[22] Daniel R Frisch. A novel technique to expand the electrocardiographic recording capability from an apple watch. *The American Journal of Medicine*, 132(8):940–941, 2019.

[23] John E Deanfield. Holter monitoring in assessment of angina pectoris. *The American Journal of Cardiology*, 59(7):C18–C22, 1987.

[24] Regina Bailey. Heart Nodes and Electrical Conduction. thoughtco. https://www.thoughtco.com/heart-nodes-anatomy-373242. Accessed: 2023-08-17.

[25] clipart library. Unlabelled diagram of the heart 1821066 (license: Personal use). https://clipart-library.com/clipart/rcnrebRMi.htm/. Accessed: 2024-04-19.

[26] Jose-Luis Cabra Lopez, Carlos Parra, Libardo Gomez, and Luis Trujillo. Sex recognition through ecg signals aiming toward smartphone authentication. *Applied Sciences*, 12(13), 2022. ISSN 2076-3417. doi: 10.3390/app12136573. URL https://www.mdpi.com/2076-3417/12/13/6573.

[27] Azure Microsoft. Artificial intelligence (AI) vs. machine learning (ML) azure. https://azure.microsoft.com/en-gb/resources/cloud-computing-dictionary/artificial-intelligence-vs-machine-learning/. Accessed: 2023-09-08.

[28] David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. A learning algorithm for boltzmann machines. *Cognitive science*, 9(1):147–169, 1985.

[29] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[30] Md Zahangir Alom, Tarek M Taha, Christopher Yakopcic, Stefan Westberg, Paheding Sidike, Mst Shamima Nasrin, Brian C Van Esesn, Abdul A S Awwal, and Vijayan K Asari. The history began from alexnet: A comprehensive survey on deep learning approaches. *arXiv preprint arXiv:1803.01164*, 2018.

[31] Noor Azwana Mat Ariff and Amelia Ritahani Ismail. Study of adam and adamax optimizers on alexnet architecture for voice biometric authentication system. In *2023 17th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, pages 1–4. IEEE, 2023.

[32] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[33] Yeong-Hyeon Byeon, Sung-Bum Pan, and Keun-Chang Kwak. Intelligent deep models based on scalograms of electrocardiogram signals for biometrics. *Sensors*, 19(4):935, 2019.

[34] Jong Min Song, Wan Kim, and Kang Ryoung Park. Finger-vein recognition based on deep densenet using composite image. *Ieee Access*, 7:66845–66863, 2019.

[35] Renuka Saravanan and Pothula Sujatha. A state of art techniques on machine learning algorithms: a perspective of supervised learning approaches in data classification. In *2018 Second international conference on intelligent computing and control systems (ICICCS)*, pages 945–949. IEEE, 2018.

[36] Xiaojin Zhu and Andrew B Goldberg. *Introduction to semi-supervised learning*. Springer Nature, 2022.

[37] Haize Hu, Jianxun Liu, Xiangping Zhang, and Mengge Fang. An effective and adaptable k-means algorithm for big data cluster analysis. *Pattern Recognition*, 139:109404, 2023.

[38] Helton Hugo de Carvalho, Robson Luiz Moreno, Tales Cleber Pimenta, Paulo C. Crepaldi, and Evaldo Cintra. A heart disease recognition embedded system with fuzzy cluster algorithm. *Computer Methods and Programs in Biomedicine*, 110(3):447–454, 2013. ISSN 0169-2607. doi: https://doi.org/10.1016/j.cmpb.2013.01.005. URL https://www.sciencedirect.com/science/article/pii/S0169260713000084.

[39] Marcelo Godoy Simões and Ian S Shaw. *Controle e modelagem fuzzy*. Editora Blucher, 2007.

[40] Yun-Chi Yeh, Wen-June Wang, and Che Wun Chiou. A novel fuzzy c-means method for classifying heartbeat cases from ecg signals. *Measurement*, 43(10):1542–1555, 2010. ISSN 0263-2241. doi: https://doi.org/10.1016/j.measurement.2010.08.019. URL https://www.sciencedirect.com/science/article/pii/S0263224110001934.

[41] Rahime Ceylan, Yüksel Özbay, and Bekir Karlik. A novel approach for classification of ecg arrhythmias: Type-2 fuzzy clustering neural network. *Expert Systems with Applications*, 36(3, Part 2):6721–6726, 2009. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2008.08.028. URL https://www.sciencedirect.com/science/article/pii/S0957417408005745.

[42] Alan Jovic and Nikola Bogunovic. Electrocardiogram analysis using a combination of statistical, geometric, and nonlinear heart rate variability features. *Artificial intelligence in medicine*, 51(3):175–186, 2011.

[43] Fahim Sufi and Ibrahim Khalil. Diagnosis of cardiovascular abnormalities from compressed ecg: a data mining-based approach. *IEEE transactions on information technology in biomedicine*, 15(1):33–39, 2010.

[44] Fahim Sufi, Ibrahim Khalil, and Abdun Naser Mahmood. A clustering based system for instant detection of cardiac abnormalities from compressed ecg. *Expert Systems with Applications*, 38(5):4705–4713, 2011.

[45] Fahim Sufi and Ibrahim Khalil. Faster person identification using compressed ecg in time critical wireless telecardiology applications. *Journal of Network and Computer Applications*, 34(1):282–293, 2011.

[46] Chuong B Do and Serafim Batzoglou. What is the expectation maximization algorithm? *Nature biotechnology*, 26(8):897–899, 2008.

[47] Selcan Kaplan Berkaya, Alper Kursat Uysal, Efnan Sora Gunal, Semih Ergin, Serkan Gunal, and M. Bilginer Gulmezoglu. A survey on ecg analysis. *Biomedical Signal Processing and Control*, 43:216–235, 2018. ISSN 1746-8094. doi: https://doi.org/10.1016/j.bspc.2018.03.003. URL https://www.sciencedirect.com/science/article/pii/S1746809418300636.

[48] Hui Yang, Chen Kan, Gang Liu, and Yun Chen. Spatiotemporal differentiation of myocardial infarctions. *IEEE Transactions on Automation Science and Engineering*, 10(4):938–947, 2013.

[49] Yun Chen and Hui Yang. Self-organized neural network for the quality control of 12-lead ecg signals. *Physiological measurement*, 33(9):1399, 2012.

[50] Willian Dihanster Gomes de Oliveira and Lilian Berton. A systematic review for class-imbalance in semi-supervised learning. *Artificial Intelligence Review*, pages 1–34, 2023.

[51] Richard Bellman. A markovian decision process. *Journal of mathematics and mechanics*, pages 679–684, 1957.

[52] Yuxi Li. Reinforcement learning applications. *arXiv preprint arXiv:1908.06973*, 2019.

[53] Aleksandr Ometov, Viktoriia Shubina, Lucie Klus, Justyna Skibińska, Salwa Saafi, Pavel Pascacio, Laura Flueratoru, Darwin Quezada Gaibor, Nadezhda Chukhno, Olga Chukhno, Asad Ali, Asma Channa, Ekaterina Svertoka, Waleed Bin Qaim, Raúl Casanova-Marqués, Sylvia Holcer, Joaquín Torres-Sospedra, Sven Casteleyn, Giuseppe Ruggeri, Giuseppe Araniti, Radim Burget, Jiri Hosek, and Elena Simona Lohan. A survey on wearable technology: History, state-of-the-art and current challenges. *Computer Networks*, 193: 108074, 2021. ISSN 1389-1286. doi: https://doi.org/10.1016/j.comnet.2021.108074. URL https://www.sciencedirect.com/science/article/pii/S1389128621001651.

[54] Daniel Maurice Lerner. User-wearable secured devices provided assuring authentication and validation of data storage and transmission, April 28 2020. US Patent 10,637,854.

[55] Christiane Attig and Thomas Franke. I track, therefore i walk–exploring the motivational costs of wearing activity trackers in actual users. *International Journal of Human-Computer Studies*, 127:211–224, 2019.

[56] Katie-Jane Brickwood, Greig Watson, Jane O'Brien, Andrew D Williams, et al. Consumer-based wearable activity trackers increase physical activity participation: systematic review and meta-analysis. *JMIR mHealth and uHealth*, 7(4):e11819, 2019.

[57] Manju Rana and Vikas Mittal. Wearable sensors for real-time kinematics analysis in sports: A review. *IEEE Sensors Journal*, 21(2):1187–1207, 2020.

[58] Sudip Vhaduri and Christian Poellabauer. Biometric-based wearable user authentication during sedentary and non-sedentary periods. *arXiv preprint arXiv:1811.07060*, 2018.

[59] Zeineb Bouzid, Salah S Al-Zaiti, Raymond Bond, and Ervin Sejdić. Remote and wearable ecg devices with diagnostic abilities in adults: A state-of-the-science scoping review. *Heart Rhythm*, 19(7):1192–1201, 2022.

[60] Dung Phan, Lee Yee Siong, Pubudu N Pathirana, and Aruna Seneviratne. Smartwatch: Performance evaluation for long-term heart rate monitoring. In *2015 International symposium on bioelectronics and bioinformatics (ISBB)*, pages 144–147. IEEE, 2015.

[61] Mohit Ingale, Renato Cordeiro, Siddartha Thentu, Younghee Park, and Nima Karimian. Ecg biometric authentication: A comparative analysis. *IEEE Access*, 8:117853–117866, 2020.

[62] Robert Rho, Mark Vossler, Susan Blancher, and Jeanne E. Poole. Comparison of 2 ambulatory patch ecg monitors: The benefit of the p-wave and signal clarity. *American Heart Journal*, 203:109–117, 2018. ISSN

0002-8703. doi: https://doi.org/10.1016/j.ahj.2018.03.022. URL https://www.sciencedirect.com/science/article/pii/S0002870318301066.

[63] Warren M. Smith, Fiona Riddell, Morag Madon, and Marye J. Gleva. Comparison of diagnostic value using a small, single channel, p-wave centric sternal ecg monitoring patch with a standard 3-lead holter system over 24 hours. *American Heart Journal*, 185:67–73, 2017. ISSN 0002-8703. doi: https://doi.org/10.1016/j.ahj.2016.11.006. URL https://www.sciencedirect.com/science/article/pii/S0002870316302733.

[64] J Muhlsteff, O Such, R Schmidt, M Perkuhn, H Reiter, J Lauter, J Thijs, G Musch, and M Harris. Wearable approach for continuous ecg-and activity patient-monitoring. In *The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, volume 1, pages 2184–2187. IEEE, 2004.

[65] Matthew Boakes. *A Performance Assessment Framework for Mobile Biometrics*. PhD thesis, University of Kent,, October 2022. URL https://kar.kent.ac.uk/97792/.

[66] Marco Santopietro. *An exploration of dynamic biometric performance using device interaction and wearable technologies*. PhD thesis, University of Kent,, November 2022. URL https://kar.kent.ac.uk/98627/.

[67] Joana S Paiva, Duarte Dias, and Joao PS Cunha. Beat-id: Towards a computationally low-cost single heartbeat biometric identity check system based on electrocardiogram wave morphology. *PloS one*, 12(7):e0180942, 2017.

[68] Vessela Krasteva, Irena Jekova, and Roger Abächerli. Biometric verification by cross-correlation analysis of 12-lead ecg patterns: Ranking of the most reliable peripheral and chest leads. *Journal of electrocardiology*, 50(6):847–854, 2017.

[69] Hyun-Soo Choi, Byunghan Lee, and Sungroh Yoon. Biometric authentication using noisy electrocardiograms acquired by mobile sensors. *IEEE access*, 4:1266–1273, 2016.

[70] Saeid Wahabi, Shahrzad Pouryayevali, Siddarth Hari, and Dimitrios Hatzinakos. On evaluating ecg biometric systems: Session-dependence and body posture. *IEEE Transactions on Information Forensics and Security*, 9(11): 2002–2013, 2014. doi: 10.1109/TIFS.2014.2360430.

[71] Nima Karimian, Zimu Guo, Mark Tehranipoor, and Domenic Forte. Highly reliable key generation from electrocardiogram (ecg). *IEEE Transactions on Biomedical Engineering*, 64(6):1400–1411, 2016.

[72] Emna Kalai Zaghouani, Adel Benzina, and Rabah Attia. Ecg based authentication for e-healthcare systems: Towards a secured ecg features transmission. In *2017 13th international wireless communications and mobile computing conference (IWCMC)*, pages 1777–1783. IEEE, 2017.

[73] Hanvit Kim, Thanh Quoc Phan, Wonjae Hong, and Se Young Chun. Physiology-based augmented deep neural network frameworks for ecg biometrics with short ecg pulses considering varying heart rates. *Pattern Recognition Letters*, 156:1–6, 2022.

[74] Florian Lehmann and Daniel Buschek. Heartbeats in the wild: a field study exploring ecg biometrics in everyday life. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.

[75] Shahrzad Pouryayevali, Saeid Wahabi, Siddarth Hari, and Dimitrios Hatzinakos. On establishing evaluation standards for ecg biometrics. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3774–3778. IEEE, 2014.

[76] Manal M Tantawi, Kenneth Revett, Abdel-Badeeh Salem, and Mohamed F Tolba. A wavelet feature extraction method for electrocardiogram (ecg)-based biometric recognition. *Signal, Image and Video Processing*, 9:1271–1280, 2015.

[77] Kiran Kumar Patro, Allam Jaya Prakash, M Jayamanmadha Rao, and P Rajesh Kumar. An efficient optimized feature selection with machine learning approach for ecg biometric recognition. *IETE Journal of Research*, 68(4): 2743–2754, 2022.

[78] Yazhao Li, Yanwei Pang, Kongqiao Wang, and Xuelong Li. Toward improving ecg biometric identification using cascaded convolutional neural networks. *Neurocomputing*, 391:83–95, 2020.

[79] Dalal A AlDuwaile and Md Saiful Islam. Using convolutional neural network and a single heartbeat for ecg biometric recognition. *Entropy*, 23(6):733, 2021.

[80] RN Begum, Ambalika Sharma, and GK Singh. Ecg based reliable user identification using deep learning. *International Journal of Biomedical and Biological Engineering*, 16(9):66–75, 2022.

[81] Mohamed Hammad, Paweł Pławiak, Kuanquan Wang, and Udyavara Rajendra Acharya. Resnet-attention model for human authentication using ecg signals. *Expert Systems*, 38(6):e12547, 2021.

[82] Basma Abd El-Rahiem and Mohamed 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*, pages 53–79, 2022.

[83] Yeong-Hyeon Byeon and Keun-Chang Kwak. Pre-configured deep convolutional neural networks with various time-frequency representations for biometrics from ecg signals. *Applied Sciences*, 9(22):4810, 2019.

[84] Teresa M. C. Pereira, Raquel C. Conceição, and Raquel Sebastião. Initial study using electrocardiogram for authentication and identification. *Sensors*, 22(6), 2022. ISSN 1424-8220. doi: 10.3390/s22062202. URL https://www.mdpi.com/1424-8220/22/6/2202.

[85] Iulian B. Ciocoiu and Nicolae Cleju. Off-person ecg biometrics using spatial representations and convolutional neural networks. *IEEE Access*, 8:218966–218981, 2020. doi: 10.1109/ACCESS.2020.3042547.

[86] Shin Jae Kang, Seung Yong Lee, Hyo Il Cho, and Hyunggon Park. Ecg authentication system design based on signal analysis in mobile and wearable devices. *IEEE Signal Processing Letters*, 23(6):805–808, 2016.

[87] Lukasz Wieclaw, Yuriy Khoma, Pawel Fałat, Dmytro Sabodashko, and Veronika Herasymenko. Biometrie identification from raw ecg signal using deep learning techniques. In *2017 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, volume 1, pages 129–133. IEEE, 2017.

[88] Eduardo Jose da Silva Luz, Gladston JP Moreira, Luiz S Oliveira, William Robson Schwartz, and David Menotti. Learning deep off-the-person heart biometrics representations. *IEEE Transactions on Information Forensics and Security*, 13(5):1258–1270, 2017.

[89] Allam Jaya Prakash, Kiran Kumar Patro, Saunak Samantray, Paweł Pławiak, and Mohamed Hammad. A deep learning technique for biometric authentication using ecg beat template matching. *Information*, 14(2):65, 2023.

[90] Carmen Camara, Pedro Peris-Lopez, Honorio Martín, and Mu'awya Aldalaien. Ecg-rng: A random number generator based on ecg signals and suitable for securing wireless sensor networks. *Sensors*, 18(9):2747, 2018.

[91] Lorena González-Manzano, José M de Fuentes, Pedro Peris-Lopez, and Carmen Camara. Encryption by heart (ebh)—using ecg for time-invariant symmetric key generation. *Future Generation Computer Systems*, 77:136–148, 2017.

[92] Saeid Wahabi, Shahrzad Pouryayevali, and Dimitrios Hatzinakos. Posture-invariant ecg recognition with posture detection. In *2015 IEEE International*

Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1812–1816, 2015. doi: 10.1109/ICASSP.2015.7178283.

[93] Jeehoon Kim, Dongsuk Sung, MyungJun Koh, Jason Kim, and Kwang Suk Park. Electrocardiogram authentication method robust to dynamic morphological conditions. *IET Biometrics*, 8(6):401–410, 2019.

[94] Juzheng Liu, Jing Chen, Hanjun Jiang, Wen Jia, Qingliang Lin, and Zhihua Wang. Activity recognition in wearable ecg monitoring aided by accelerometer data. In *2018 IEEE international symposium on circuits and systems (ISCAS)*, pages 1–4. IEEE, 2018.

[95] Sakorn Mekruksavanich and Anuchit Jitpattanakul. Biometric user identification based on human activity recognition using wearable sensors: An experiment using deep learning models. *Electronics*, 10(3):308, 2021.

[96] Samera Batool, Nazar A Saqib, and Muazzam A Khan. Internet of things data analytics for user authentication and activity recognition. In *2017 Second International Conference on Fog and Mobile Edge Computing (FMEC)*, pages 183–187. IEEE, 2017.

[97] Fatima Sajid Butt, Luigi La Blunda, Matthias F. Wagner, Jörg Schäfer, Inmaculada Medina-Bulo, and David Gómez-Ullate. Fall detection from electrocardiogram (ecg) signals and classification by deep transfer learning. *Information*, 12(2), 2021. ISSN 2078-2489. doi: 10.3390/info12020063. URL https://www.mdpi.com/2078-2489/12/2/63.

[98] Gloria Cosoli, Luca Antognoli, and Lorenzo Scalise. Wearable electrocardiography for physical activity monitoring: Definition of validation protocol and automatic classification. *Biosensors*, 13(2):154, 2023.

[99] Paloma Tirado-Martin, Judith Liu-Jimenez, Jorge Sanchez-Casanova, and Raul Sanchez-Reillo. Qrs differentiation to improve ecg biometrics under different physical scenarios using multilayer perceptron. *Applied Sciences*, 10(19):6896, 2020.

[100] Muhammad Muizz Mohd Nawawi, Khairul Azami Sidek, and Amelia Wong Azman. Ecg biometric in real-life settings: analysing different physiological conditions with wearable smart textiles shirts. *Bulletin of Electrical Engineering and Informatics*, 12(5):2930–2938, 2023.

[101] Susana Brás, Jacqueline HT Ferreira, Sandra C Soares, and Armando J Pinho. Biometric and emotion identification: An ecg compression based method. *Frontiers in psychology*, 9:467, 2018.

[102] Wei Li, Cheng Fang, Zhihao Zhu, Chuyi Chen, and Aiguo Song. Turning waste into wealth: Person identification by emotion-disturbed electrocardiogram. *IET Biometrics*, 12(3):159–175, 2023.

[103] Shivangi Gupta and Sunil Kumar Chowdhary. Authentication through electrocardiogram signals based on emotions-a step towards atm security. In *2017 6th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*, pages 440–442. IEEE, 2017.

[104] Mariana Vaz, Teresa Summavielle, Raquel Sebastião, and Rita P. Ribeiro. Multimodal classification of anxiety based on physiological signals. *Applied Sciences*, 13(11), 2023. ISSN 2076-3417. doi: 10.3390/app13116368. URL https://www.mdpi.com/2076-3417/13/11/6368.

[105] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM international conference on multimodal interaction*, pages 400–408, 2018.

[106] Prerna Garg, Jayasankar Santhosh, Andreas Dengel, and Shoya Ishimaru. Stress detection by machine learning and wearable sensors. In *26th International Conference on Intelligent User Interfaces-Companion*, pages 43–45, 2021.

[107] Jionghao Lin, Shirui Pan, Cheng Siong Lee, and Sharon Oviatt. An explainable deep fusion network for affect recognition using physiological signals. In

Proceedings of the 28th ACM International Conference on Information and Knowledge Management, pages 2069–2072, 2019.

[108] Pritam Sarkar and Ali Etemad. Self-supervised ecg representation learning for emotion recognition. *IEEE Transactions on Affective Computing*, 2020.

[109] Behnam Behinaein, Anubhav Bhatti, Dirk Rodenburg, Paul Hungler, and Ali Etemad. A transformer architecture for stress detection from ecg. In *2021 International Symposium on Wearable Computers*, pages 132–134, 2021.

[110] Ritu Tanwar, Orchid Chetia Phukan, Ghanapriya Singh, and Sanju Tiwari. Cnn-lstm based stress recognition using wearables. 2022.

[111] Adrian Colomer Granero, Félix Fuentes-Hurtado, Valery Naranjo Ornedo, Jaime Guixeres Provinciale, Jose M Ausín, and Mariano Alcaniz Raya. A comparison of physiological signal analysis techniques and classifiers for automatic emotional evaluation of audiovisual contents. *Frontiers in computational neuroscience*, 10:74, 2016.

[112] Wonki Lee, Seulgee Kim, and Daeeun Kim. Individual biometric identification using multi-cycle electrocardiographic waveform patterns. *Sensors*, 18(4), 2018. ISSN 1424-8220. doi: 10.3390/s18041005. URL https://www.mdpi.com/1424-8220/18/4/1005.

[113] Gilberto Perpiñan, Erika Severeyn, Miguel Altuve, and Sara Wong. Classification of metabolic syndrome subjects and marathon runners with the k-means algorithm using heart rate variability features. In *2016 XXI Symposium on Signal Processing, Images and Artificial Vision (STSIVA)*, pages 1–6. IEEE, 2016.

[114] BW Johansson. [a history of the electrocardiogram]. *Dansk medicinhistorisk arbog*, page 163—176, 2001. ISSN 0084-9588. URL http://europepmc.org/abstract/MED/11848076.

[115] CJ Wiggers, R Wegria, and B Pinera. The effects of myocardial ischemia on the fibrillation threshold—the mechanism of spontaneous ventricular fibrillation following coronary occlusion. *American Journal of Physiology-Legacy Content*, 131(2):309–316, 1940.

[116] NEIL DE SOYZA, JOE K. BISSETT, JAMES J. KANE, MARVIN L. MURPHY, and JAMES E. DOHERTY. Ectopic ventricular prematurity and its relationship to ventricular tachycardia in acute myocardial infarction in man. *Circulation*, 50(3):529–533, 1974. doi: 10.1161/01.CIR.50.3.529. URL https://www.ahajournals.org/doi/abs/10.1161/01.CIR.50.3.529.

[117] William C. Roberts and Marc A. Silver. Norman jefferis holter and ambulatory ecg monitoring. *The American Journal of Cardiology*, 52(7):903–906, 1983. ISSN 0002-9149. doi: https://doi.org/10.1016/0002-9149(83)90439-3. URL https://www.sciencedirect.com/science/article/pii/0002914983904393.

[118] David Giles, Nick Draper, and William Neil. Validity of the polar v800 heart rate monitor to measure rr intervals at rest. *European journal of applied physiology*, 116:563–571, 2016.

[119] Vega Pradana Rachim and Wan-Young Chung. Wearable noncontact armband for mobile ecg monitoring system. *IEEE Transactions on Biomedical Circuits and Systems*, 10(6):1112–1118, 2016. doi: 10.1109/TBCAS.2016.2519523.

[120] Nino Isakadze and Seth S Martin. How useful is the smartwatch ecg? *Trends in cardiovascular medicine*, 30(7):442–448, 2020.

[121] Brinnae Bent, Benjamin A Goldstein, Warren A Kibbe, and Jessilyn P Dunn. Investigating sources of inaccuracy in wearable optical heart rate sensors. *NPJ digital medicine*, 3(1):18, 2020.

[122] Awni Y Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H Tison, Codie Bourn, Mintu P Turakhia, and Andrew Y Ng. Cardiologist-level

arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature medicine*, 25(1):65–69, 2019.

[123] George B Moody and Roger G Mark. Development and evaluation of a 2-lead ecg analysis program. *Computers in cardiology*, 9:39–44, 1982.

[124] Zachi I Attia, Peter A Noseworthy, Francisco Lopez-Jimenez, Samuel J Asirvatham, Abhishek J Deshmukh, Bernard J Gersh, Rickey E Carter, Xiaoxi Yao, Alejandro A Rabinstein, Brad J Erickson, et al. An artificial intelligence-enabled ecg algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *The Lancet*, 394(10201):861–867, 2019.

[125] Jia-wei Zhang, Li-ping Wang, Xia Liu, Hong-hai Zhu, and Jun Dong. Chinese cardiovascular disease database (ccdd) and its management tool. In *2010 IEEE International Conference on BioInformatics and BioEngineering*, pages 66–72. IEEE, 2010.

[126] Linpeng Jin and Jun Dong. Classification of normal and abnormal ecg records using lead convolutional neural network and rule inference. *Science China Information Sciences*, 60:1–3, 2017.

[127] Jean-Philippe Couderc. The telemetric and holter ecg warehouse initiative (thew): a data repository for the design, implementation and validation of ecg-related technologies. In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, pages 6252–6255. IEEE, 2010.

[128] MA García-González, A Argelagós, Mireya Fernández-Chimeno, and J Ramos-Castro. Differences in qrs locations due to ecg lead: relationship with breathing. In *XIII Mediterranean Conference on Medical and Biological Engineering and Computing 2013: MEDICON 2013, 25-28 September 2013, Seville, Spain*, pages 962–964. Springer, 2014.

[129] Miguel A García-González, Ariadna Argelagós-Palau, Mireya Fernández-Chimeno, and Juan Ramos-Castro. A comparison of heartbeat detectors

for the seismocardiogram. In *Computing in Cardiology 2013*, pages 461–464. IEEE, 2013.

[130] Qingxue Zhang, Dian Zhou, and Xuan Zeng. A novel framework for motion-tolerant instantaneous heart rate estimation by phase-domain multiview dynamic time warping. *IEEE Transactions on Biomedical Engineering*, 64(11): 2562–2574, 2016.

[131] JL Willems, P Arnaud, JH Van Bemmel, R Degani, PW Macfarlane, Chr Zywietz, et al. Common standards for quantitative electrocardiography: goals and main results. *Methods of information in medicine*, 29(04):263–271, 1990.

[132] Radovan Smíšek, Lucie Maršánová, Andrea Němcová, Martin Vítek, Jiří Kozumplík, and Marie Nováková. Cse database: extended annotations and new recommendations for ecg software testing. *Medical & Biological Engineering & Computing*, 55:1473–1482, 2017.

[133] Ralf Bousseljot, Dieter Kreiseler, and Allard Schnabel. Nutzung der ekg-signaldatenbank cardiodat der ptb über das internet. 1995.

[134] A Taddei, G Distante, M Emdin, P Pisani, GB Moody, C Zeelenberg, and C Marchesi. The european st-t database: standard for evaluating systems for the analysis of st-t changes in ambulatory electrocardiography. *European heart journal*, 13(9):1164–1172, 1992.

[135] Gari D Clifford, Chengyu Liu, Benjamin Moody, H Lehman Li-wei, Ikaro Silva, Qiao Li, AE Johnson, and Roger G Mark. Af classification from a short single lead ecg recording: The physionet/computing in cardiology challenge 2017. In *2017 Computing in Cardiology (CinC)*, pages 1–4. IEEE, 2017.

[136] Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I Lunze, Wojciech Samek, and Tobias Schaeffter. Ptb-xl, a large publicly available electrocardiography dataset. *Scientific data*, 7(1):154, 2020.

[137] Tatiana S Lugovaya. Biometric human identification based on electrocardiogram. *Master's thesis, Faculty of Computing Technologies and Informatics, Electrotechnical University 'LETI', Saint-Petersburg, Russian Federation*, 2005.

[138] Jean-Philippe Couderc, Xia Xiaojuan, Wojciech Zareba, and Arthur J Moss. Assessment of the stability of the individual-based correction of qt interval for heart rate. *Annals of noninvasive electrocardiology : the official journal of the International Society for Holter and Noninvasive Electrocardiology, Inc*, 10(1):25–34, 2005.

[139] M Vollmer, D Bläsing, M Nisser, A Buder, et al. Simultaneous physiological measurements with five devices at different cognitive and physical loads (version 1.0. 0). *PhysioNet*, 2020.

[140] George B Moody and Roger G Mark. The impact of the mit-bih arrhythmia database. *IEEE engineering in medicine and biology magazine*, 20(3):45–50, 2001.

[141] George B Moody, W Muldrow, and Roger G Mark. A noise stress test for arrhythmia detectors. *Computers in cardiology*, 11(3):381–384, 1984.

[142] FM Nolle, FK Badura, JM Catlett, RW Bowser, and MH Sketch. Crei-gard, a new concept in computerized arrhythmia monitoring systems. *Computers in Cardiology*, 13(1):515–518, 1986.

[143] Scott David Greenwald. *The development and analysis of a ventricular fibrillation detector*. PhD thesis, Massachusetts Institute of Technology, 1986.

[144] Oresti Banos, Rafael Garcia, Juan A Holgado-Terriza, Miguel Damas, Hector Pomares, Ignacio Rojas, Alejandro Saez, and Claudia Villalonga. mhealthdroid: a novel framework for agile development of mobile health applications. In *Ambient Assisted Living and Daily Activities: 6th International Work-Conference, IWAAL 2014, Belfast, UK, December 2-5, 2014. Proceedings 6*, pages 91–98. Springer, 2014.

[145] Vikto Tihonenko, A Khaustov, S Ivanov, A Rivin, and E Yakushenko. St petersburg incart 12-lead arrhythmia database. *PhysioBank PhysioToolkit and PhysioNet*, 2008.

[146] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

[147] Jianwei Zheng, Jianming Zhang, Sidy Danioko, Hai Yao, Hangyuan Guo, and Cyril Rakovski. A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. *Scientific data*, 7(1):48, 2020.

[148] Nikhil Iyengar, CK Peng, Raymond Morin, Ary L Goldberger, and Lewis A Lipsitz. Age-related alterations in the fractal scaling of cardiac interbeat interval dynamics. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, 271(4):R1078–R1084, 1996.

[149] Simona Petrutiu, Alan V Sahakian, and Steven Swiryn. Abrupt changes in fibrillatory wave characteristics at the termination of paroxysmal atrial fibrillation in humans. *Europace*, 9(7):466–470, 2007.

[150] Hugo Plácido Da Silva, André Lourenço, Ana Fred, Nuno Raposo, and Marta Aires-de Sousa. Check your biosignals here: A new dataset for off-the-person ecg biometrics. *Computer methods and programs in biomedicine*, 113(2):503–514, 2014.

[151] Juan Pablo Martínez, Olle Pahlm, Michael Ringborn, Stafford Warren, Pablo Laguna, and Leif Sörnmo. The staff iii database: Ecgs recorded during acutely induced myocardial ischemia. In *2017 Computing in Cardiology (CinC)*, pages 1–4. IEEE, 2017.

[152] Feifei Liu, Chengyu Liu, Lina Zhao, Xiangyu Zhang, Xiaoling Wu, Xiaoyan Xu, Yulin Liu, Caiyun Ma, Shoushui Wei, Zhiqiang He, et al. An open

access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. *Journal of Medical Imaging and Health Informatics*, 8(7):1368–1373, 2018.

[153] ECG and ECHO Learning the ecg leads: Electrodes,limb leads, chest (precordial) leads and 12 leads ecg. https://ecgwaves.com/topic/ekg-ecg-leads-electrodes-systems-limb-chest-precordial/, . Accessed: 2024-05-08.

[154] Ruggero Donida Labati, Enrique Muñoz, Vincenzo Piuri, Roberto Sassi, and Fabio Scotti. Deep-ecg: Convolutional neural networks for ecg biometric recognition. *Pattern Recognition Letters*, 126:78–85, 2019. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2018.03.028. URL https://www.sciencedirect.com/science/article/pii/S0167865518301077. Robustness, Security and Regulation Aspects in Current Biometric Systems.

[155] UCI Machine Learning Repository center for machine learning and intelligent systems. https://archive.ics.uci.edu/ml/datasets/WESAD+%28Wearable+Stress+and+Affect+Detection%29/. Accessed: 2023-02-05.

[156] Plux Wireless Biosignals biosignal plux respiban professional. https://support.pluxbiosignals.com/knowledge-base/is-biosignalsplux-a-medical-device/. Accessed: 2023-02-06.

[157] Empetica INC e4 wristband. https://www.empatica.com/en-gb/research/e4/. Accessed: 2023-02-06.

[158] Andrea C Samson, Sylvia D Kreibig, Blake Soderstrom, A Ayanna Wade, and James J Gross. Eliciting positive, negative and mixed emotional states: A film library for affective scientists. *Cognition and emotion*, 30(5):827–856, 2016.

[159] Yanmin Sun, Andrew KC Wong, and Mohamed S Kamel. Classification of imbalanced data: A review. *International journal of pattern recognition and artificial intelligence*, 23(04):687–719, 2009.

[160] Philip Okoampah Kwaning. Techniques for Handling Imbalanced Datasets in Machine Learning medium. https://medium.com/@philipokoampah/techniques-for-handling-imbalanced-datasets-in-machine-learning\-7e560907af30/. Accessed: 2023-05-09.

[161] Satyendra Singh Rawat and Amit Kumar Mishra. Review of methods for handling class-imbalanced in classification problems. *arXiv preprint arXiv:2211.05456*, 2022.

[162] Mind Media Neuro and Biofeedback Systems nexus-10 mkii. https://www.mindmedia.com/en/products/nexus-10-mkii/. Accessed: 2023-02-01.

[163] Bittium Corporation bittium emotion faros 360°. https://shop.bittium.com/product/37/emotion-faros-360. Accessed: 2023-02-01.

[164] Grzegorz Bilo, Cristina Zorzi, Juan E Ochoa Munera, Camilla Torlasco, Valentina Giuli, and Gianfranco Parati. Validation of the somnotouch-nibp noninvasive continuous blood pressure monitor according to the european society of hypertension international protocol revision 2010. *Blood pressure monitoring*, 20(5):291, 2015.

[165] Hexoskin Health Sensors and AI hexoskin. https://www.hexoskin.com/pages/health-research. Accessed: 2023-02-01.

[166] Ruggero Donida Labati, Roberto Sassi, and Fabio Scotti. Ecg biometric recognition: Permanence analysis of qrs signals for 24 hours continuous authentication. In *2013 IEEE international workshop on information forensics and security (WIFS)*, pages 31–36. IEEE, 2013.

[167] Behnam Behinaein, Anubhav Bhatti, Dirk Rodenburg, Paul Hungler, and Ali Etemad. A transformer architecture for stress detection from ecg. In *Proceedings of the 2021 ACM International Symposium on Wearable Computers*, ISWC '21, page 132–134, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384629. doi: 10.1145/3460421.3480427. URL https://doi.org/10.1145/3460421.3480427.

[168] Dominique Makowski, Tam Pham, Zen J Lau, Jan C Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and SH Annabel Chen. Neurokit2: A python toolbox for neurophysiological signal processing. *Behavior research methods*, pages 1–8, 2021.

[169] Zeeshan Hassan, Syed Omer Gilani, and Mohsin Jamil. Review of fiducial and non-fiducial techniques of feature extraction in ecg based biometric systems. *Indian J. Sci. Technol*, 9(21):850–855, 2016.

[170] Manal M Tantawi, Kenneth Revett, A Salem, and Mohamed Fahmy Tolba. Fiducial feature reduction analysis for electrocardiogram (ecg) based biometric recognition. *Journal of Intelligent Information Systems*, 40:17–39, 2013.

[171] Sandipan Dhar, Abhishek Chakraborty, Deboleena Sadhukhan, Saurabh Pal, and Madhuchhanda Mitra. Efficient detection of cardiac abnormalities via a simplified score-based analysis of the ecg signal. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–12, 2024.

[172] Jiapu Pan and Willis J Tompkins. A real-time qrs detection algorithm. *IEEE transactions on biomedical engineering*, (3):230–236, 1985.

[173] ECG and ECHO Learning ecg interpretation: Characteristics of the normal ecg (p-wave, qrs complex, st segment, t-wave). https://ecgwaves.com/topic/ecg-normal-p-wave-qrs-complex-st-segment-t-wave-j-point/, . Accessed: 2023-04-25.

[174] Yogendra Narain Singh, Sanjay Kumar Singh, and Amit Kumar Ray. Bioelectrical signals as emerging biometrics: Issues and challenges. *International Scholarly Research Notices*, 2012, 2012.

[175] E. S. Prakash and Madanmohan. How to tell heart rate from an ecg? (learning objects #769 and #878). *Advances in Physiology Education*, 29(2):57–57, 2005. doi: 10.1152/advan.00013.2005.

[176] Deboleena Sadhukhan and Madhuchhanda Mitra. R-peak detection algorithm for ecg using double difference and rr interval processing. *Procedia Technology*, 4:873–877, 2012. ISSN 2212-0173. doi: https://doi.org/10.1016/j.protcy.2012.05.143. URL https://www.sciencedirect.com/science/article/pii/S2212017312004227. 2nd International Conference on Computer, Communication, Control and Information Technology(C3IT-2012) on February 25 - 26, 2012.

[177] Hirofumi Tanaka, Kevin D Monahan, and Douglas R Seals. Age-predicted maximal heart rate revisited. *Journal of the American College of Cardiology*, 37(1):153–156, 2001. doi: 10.1016/S0735-1097(00)01054-8.

[178] Nurul Ashikin Abdul-Kadir, Norlaili Mat Safri, and Mohd Afzan Othman. Dynamic ecg features for atrial fibrillation recognition. *Computer methods and programs in biomedicine*, 136:143–150, 2016.

[179] Hooman Sedghamiz. An online algorithm for r, s and t wave detection. *Matlab Central Community Profile, doi*, 10, 2013.

[180] What is the k-nearest neighbors algorithm? ibm. https://www.ibm.com/topics/knn, . Accessed: 2023-02-18.

[181] Taxicab geometry wikipedia. https://en.wikipedia.org/wiki/Taxicab_geometry. Accessed: 2023-02-27.

[182] Towards Data Science 9 distance measures in data science. https://towardsdatascience.com/9-distance-measures-in-data-science-918109d069fa. Accessed: 2023-05-10.

[183] Martin Zihlmann, Dmytro Perekrestenko, and Michael Tschannen. Convolutional recurrent neural networks for electrocardiogram classification. In *2017 Computing in Cardiology (CinC)*, pages 1–4. IEEE, 2017.

[184] Hamid Khorrami and Majid Moavenian. A comparative study of dwt, cwt and dct transformations in ecg arrhythmias classification. *Expert systems with Applications*, 37(8):5751–5757, 2010.

[185] Md Shamim Towhid and Md Mijanur Rahman. Spectrogram segmentation for bird species classification based on temporal continuity. In *2017 20th International Conference of Computer and Information Technology (ICCIT)*, pages 1–4. IEEE, 2017.

[186] Hao Meng, Tianhao Yan, Fei Yuan, and Hongwei Wei. Speech emotion recognition from 3d log-mel spectrograms with deep learning network. *IEEE access*, 7:125868–125881, 2019.

[187] Mike Cadogan and Robert Buttner. Ecg rate interpretation.

[188] Jonathan W Stubblefield. *Artificial Intelligence Algorithms for Medical Imaging and Healthcare*. Arkansas State University, 2021.

[189] Mario Merone, Paolo Soda, Mario Sansone, and Carlo Sansone. Ecg databases for biometric systems: A systematic review. *Expert Systems with Applications*, 67:189–202, 2017. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2016.09.030. URL https://www.sciencedirect.com/science/article/pii/S0957417416305140.

[190] Nuno Bento, David Belo, and Hugo Gamboa. Ecg biometrics using spectrograms and deep neural networks. *Int. J. Mach. Learn. Comput*, 10(2): 259–264, 2019.

[191] Bhekumuzi M Mathunjwa, Yin-Tsong Lin, Chien-Hung Lin, Maysam F Abbod, and Jiann-Shing Shieh. Ecg arrhythmia classification by using a recurrence plot and convolutional neural network. *Biomedical Signal Processing and Control*, 64:102262, 2021.

[192] Liudmila A Manilo, Anatoly P Nemirko, Ekaterina G Evdakova, and Anna A Tatarinova. Ecg database for evaluating the efficiency of recognizing dangerous arrhythmias. In *2021 IEEE Ural-Siberian Conference on Computational*

Technologies in Cognitive Science, Genomics and Biomedicine (CSGB), pages 120–123. IEEE, 2021.

[193] Qingxue Zhang, Dian Zhou, and Xuan Zeng. Heartid: A multiresolution convolutional neural network for ecg-based biometric human identification in smart health applications. *Ieee Access*, 5:11805–11816, 2017.

[194] Sara S Abdeldayem and Thirimachos Bourlai. A novel approach for ecg-based human identification using spectral correlation and deep learning. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2(1): 1–14, 2019.

[195] Yeong-Hyeon Byeon, Sung-Bum Pan, and Keun-Chang Kwak. Ensemble deep learning models for ecg-based biometrics. In *2020 Cybernetics Informatics (KI)*, pages 1–5, 2020. doi: 10.1109/KI48306.2020.9039871.

[196] Yingjie Li and Gang Liu. Sound classification based on spectrogram for surveillance applications. In *2016 IEEE International Conference on Network Infrastructure and Digital Content (IC-NIDC)*, pages 293–297, 2016. doi: 10.1109/ICNIDC.2016.7974583.

[197] W Koenig, HK Dunn, and LY Lacy. The sound spectrograph. *The Journal of the Acoustical Society of America*, 18(1):19–49, 1946.

[198] V. Zue and R. Cole. Experiments on spectrogram reading. In *ICASSP '79. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages 116–119, 1979. doi: 10.1109/ICASSP.1979.1170735.

[199] Boualem Boashash. *Time-frequency signal analysis and processing: a comprehensive reference*. Academic press, 2015.

[200] Hongzu Li and Pierre Boulanger. Structural anomalies detection from electrocardiogram (ecg) with spectrogram and handcrafted features. *Sensors*, 22(7), 2022. ISSN 1424-8220. doi: 10.3390/s22072467. URL https://www.mdpi.com/1424-8220/22/7/2467.

[201] Stanley S Stevens. The relation of pitch to intensity. *The Journal of the Acoustical Society of America*, 6(3):150–154, 1935.

[202] Stanley Smith Stevens, John Volkmann, and Edwin Broomell Newman. A scale for the measurement of the psychological magnitude pitch. *The journal of the acoustical society of america*, 8(3):185–190, 1937.

[203] William 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. *Journal of the International Phonetic Association*, 20(2):52–54, 1990. doi: 10.1017/S002510030000431X.

[204] Hao Meng, Tianhao Yan, Fei Yuan, and Hongwei Wei. Speech emotion recognition from 3d log-mel spectrograms with deep learning network. *IEEE Access*, 7:125868–125881, 2019. doi: 10.1109/ACCESS.2019.2938007.

[205] Mel Spectrogram mathworks. https://uk.mathworks.com/help/audio/ref/melspectrogram.html, . Accessed: 2023-05-07.

[206] Anfal Ahmed Aleidan, Qaisar Abbas, Yassine Daadaa, Imran Qureshi, Ganeshkumar Perumal, Mostafa EA Ibrahim, and Alaa ES Ahmed. Biometric-based human identification using ensemble-based technique and ecg signals. *Applied Sciences*, 13(16):9454, 2023.

[207] Scalogram Computation in Signal Analyzer mathworks. https://uk.mathworks.com/help/signal/ug/scalogram-computation-in-signal-analyzer.html, . Accessed: 2023-03-02.

[208] S Mallet. A wavelet tour of signal processing, 2nd edn, 1999.

[209] Continuous wavelet transform filter bank mathworks. https://uk.mathworks.com/help/wavelet/ref/cwtfilterbank.html, . Accessed: 2023-03-02.

[210] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, and Jonathan Taylor. *An introduction to statistical learning: With applications in python*. Springer Nature, 2023.

[211] Merve Veziroğlu, Erkan Veziroğlu, and İhsan Ömür Bucak. Performance comparison between naive bayes and machine learning algorithms for news classification. In İhsan Ömür Bucak, editor, *Bayesian Inference*, chapter 5. IntechOpen, Rijeka, 2024. doi: 10.5772/intechopen.1002778. URL https://doi.org/10.5772/intechopen.1002778.

[212] Kirtika Yadav and Reema Thareja. Comparing the performance of naive bayes and decision tree classification using r. *International Journal of Intelligent Systems and Applications*, 11(12):11–19, 2019.

[213] Shahadat Uddin, Arif Khan, Md Ekramul Hossain, and Mohammad Ali Moni. Comparing different supervised machine learning algorithms for disease prediction. *BMC medical informatics and decision making*, 19(1):1–16, 2019.

[214] Bartu Yesilkaya, Matjaž Perc, and Yalcin Isler. Manifold learning methods for the diagnosis of ovarian cancer. *Journal of Computational Science*, 63: 101775, 2022.

[215] Richard O Duda, Peter E Hart, et al. *Pattern classification and scene analysis*, volume 3. Wiley New York, 1973.

[216] Tin Kam Ho. Nearest neighbors in random subspaces. In Adnan Amin, Dov Dori, Pavel Pudil, and Herbert Freeman, editors, *Advances in Pattern Recognition*, pages 640–648, Berlin, Heidelberg, 1998. Springer Berlin Heidelberg. ISBN 978-3-540-68526-5.

[217] Stanislaw Osowski, Krzysztof Siwek, and Tomasz Markiewicz. Mlp and svm networks-a comparative study. In *Proceedings of the 6th Nordic Signal Processing Symposium, 2004. NORSIG 2004.*, pages 37–40. IEEE, 2004.

[218] Philip H Swain and Hans Hauska. The decision tree classifier: Design and potential. *IEEE Transactions on Geoscience Electronics*, 15(3):142–147, 1977.

[219] What is a Decision Tree? ibm. https://www.ibm.com/topics/decision-trees, . Accessed: 2023-02-19.

[220] Leo Breiman. Bagging predictors. *Machine learning*, 24:123–140, 1996.

[221] Thomas G Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems: First International Workshop, MCS 2000 Cagliari, Italy, June 21–23, 2000 Proceedings 1*, pages 1–15. Springer, 2000.

[222] Mahreen Ahmed, Hammad Afzal, Imran Siddiqi, Muhammad Faisal Amjad, and Khawar Khurshid. Exploring nested ensemble learners using overproduction and choose approach for churn prediction in telecom industry. *Neural Computing and Applications*, 32:3237–3251, 2020.

[223] Binh Tran, Bing Xue, and Mengjie Zhang. Genetic programming for multiple-feature construction on high-dimensional classification. *Pattern Recognition*, 93:404–417, 2019. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2019.05.006. URL https://www.sciencedirect.com/science/article/pii/S0031320319301815.

[224] Jacob Solawetz. Roboflow train, validation, test split for machine learning. https://blog.roboflow.com/train-test-split/. Accessed: 2023-05-30.

[225] Rukshan Pramoditha. Towerd Data Science why do we need a validation set in addition to training and test sets? https://towardsdatascience.com/why-do-we-need-a-validation-set-in-addition-to-training-and-test-sets-5cf4a65550e0. Accessed: 2023-05-30.

[226] Chaoning Zhang, Philipp Benz, Dawit Mureja Argaw, Seokju Lee, Junsik Kim, Francois Rameau, Jean-Charles Bazin, and In So Kweon. Resnet or densenet? introducing dense shortcuts to resnet. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3550–3559, 2021.

[227] Matlab MathWorks options for training deep learning neural network. https://uk.mathworks.com/help/deeplearning/ref/trainingoptions.html. Accessed: 2023-05-29.

[228] Ibrahem Kandel and Mauro Castelli. The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset. *ICT express*, 6(4):312–315, 2020.

[229] Ching-Ta Lu, Chun-Jen Ou, and Yen-Yu Lu. A practical app for quickly calculating the number of people using machine learning and convolutional neural networks. *Applied Sciences*, 12(12):6239, 2022.

[230] Sunitha Basodi, Chunyan Ji, Haiping Zhang, and Yi Pan. Gradient amplification: An efficient way to train deep neural networks. *Big Data Mining and Analytics*, 3(3):196–207, 2020. doi: 10.26599/BDMA.2020.9020004.

[231] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[232] Fully Connected Layers in Convolutional Neural Networks indian tech warrior machine learning. https://indiantechwarrior.com/fully-connected-layers-in-convolutional-neural-networks/. Accessed: 2023-06-13.

[233] Batch Normalization in Convolutional Neural Networks martin riva. https://www.baeldung.com/cs/batch-normalization-cnn. Accessed: 2023-06-13.

[234] Gian Carlo Cardarilli, Luca Di Nunzio, Rocco Fazzolari, Daniele Giardino, Alberto Nannarelli, Marco Re, and Sergio Spanò. A pseudo-softmax function for hardware-based high speed image classification. *Scientific reports*, 11(1):15307, 2021.

[235] MatLab MathWorks directed acyclic graph (dag) network for deep learning. https://uk.mathworks.com/help/deeplearning/ref/dagnetwork.html, . Accessed: 2023-05-30.

[236] 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: 2023-05-30.

[237] Laith Alzubaidi, Jinglan Zhang, Amjad J Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, José Santamaría, Mohammed A Fadhel, Muthana Al-Amidie, and Laith Farhan. Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions. *Journal of big Data*, 8:1–74, 2021.

[238] Mei Chee Leong, Dilip K Prasad, Yong Tsui Lee, and Feng Lin. Semi-cnn architecture for effective spatio-temporal learning in action recognition. *Applied Sciences*, 10(2):557, 2020.

[239] Dominic Bläsing, Anja Buder, Julian Elias Reiser, Maria Nisser, Steffen Derlien, and Marcus Vollmer. Ecg performance in simultaneous recordings of five wearable devices using a new morphological noise-to-signal index and smith-waterman-based rr interval comparisons. *Plos one*, 17(10):e0274994, 2022.

[240] Sena Yağmur ŞEN and Nalan ÖZKURT. Convolutional neural network hyperparameter tuning with adam optimizer for ecg classification. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 1–6. IEEE, 2020.

[241] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine*

Learning Research, pages 1139–1147, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/sutskever13.html.

[242] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[243] Shiv Ram Dubey, Satish Kumar Singh, and Bidyut Baran Chaudhuri. Adanorm: Adaptive gradient norm correction based optimizer for cnns. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5284–5293, January 2023.

[244] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.

[245] Xiang Yu, Shui-Hua Wang, Xin Zhang, and Yu-Dong Zhang. Detection of covid-19 by googlenet-cod. In De-Shuang Huang, Vitoantonio Bevilacqua, and Abir Hussain, editors, *Intelligent Computing Theories and Application*, pages 499–509, Cham, 2020. Springer International Publishing.

[246] Nima Tajbakhsh, Jae Y Shin, Suryakanth R Gurudu, R Todd Hurst, Christopher B Kendall, Michael B Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, 2016.

[247] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. *Advances in neural information processing systems*, 32, 2019.

[248] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655. PMLR, 2014.

[249] Olena Pavliuk, Myroslav Mishchuk, and Christine Strauss. Transfer learning approach for human activity recognition based on continuous wavelet transform. *Algorithms*, 16(2), 2023. ISSN 1999-4893. doi: 10.3390/a16020077. URL https://www.mdpi.com/1999-4893/16/2/77.

[250] Hao Du, Yuan He, and Tian Jin. Transfer learning for human activities classification using micro-doppler spectrograms. In *2018 IEEE International Conference on Computational Electromagnetics (ICCEM)*, pages 1–3, 2018. doi: 10.1109/COMPEM.2018.8496654.

[251] Jerritta Selvaraj, Murugappan Murugappan, Khairunizam Wan, and Sazali Yaacob. Classification of emotional states from electrocardiogram signals: a non-linear approach based on hurst. *Biomedical engineering online*, 12(1): 1–18, 2013.

[252] Ming Li, Viktor Rozgić, Gautam Thatte, Sangwon Lee, Adar Emken, Murali Annavaram, Urbashi Mitra, Donna Spruijt-Metz, and Shrikanth Narayanan. Multimodal physical activity recognition by fusing temporal and cepstral information. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 18(4):369–380, 2010.

[253] Omair Rashed Abdulwareth Almanifi, Ismail Mohd Khairuddin, Mohd Azraai Mohd Razman, Rabiu Muazu Musa, and Anwar P.P. Abdul Majeed. Human activity recognition based on wrist ppg via the ensemble method. *ICT Express*, 8(4):513–517, 2022. ISSN 2405-9595. doi: https://doi.org/10.1016/j.icte.2022.03.006. URL https://www.sciencedirect.com/science/article/pii/S2405959522000479.

[254] Justin Gilmore and Mona Nasseri. Human activity recognition algorithm with physiological and inertial signals fusion: Photoplethysmography, electrodermal activity, and accelerometry. *Sensors*, 24(10), 2024. ISSN 1424-8220. doi: 10.3390/s24103005. URL https://www.mdpi.com/1424-8220/24/10/3005.

[255] Paloma Tirado-Martin and Raul Sanchez-Reillo. Bioecg: Improving ecg biometrics with deep learning and enhanced datasets. *Applied Sciences*, 11 (13):5880, 2021.

[256] Ruggero Donida Labati, Vincenzo Piuri, Roberto Sassi, and Fabio Scotti. Heartcode: A novel binary ecg-based template. In *2014 IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications (BIOMS) Proceedings*, pages 86–91. IEEE, 2014.

[257] Shiva Asadianfam, Mohammad Javad Talebi, and Elaheh Nikougoftar. Ecg-based authentication systems: a comprehensive and systematic review. *Multimedia Tools and Applications*, pages 1–55, 2023.