IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 5 October 2022, accepted 31 October 2022, date of publication 7 November 2022, date of current version 15 November 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3220670

APPLIED RESEARCH

Toward Automated Feature Extraction for Deep Learning Classification of Electrocardiogram Signals

FATIMA SAJID BUTT[1,2], MATTHIAS F. WAGNER[1], (Member, IEEE), JÖRG SCHÄFER[1], (Member, IEEE), AND DAVID GOMEZ ULLATE[2,3]

[1]Faculty 2 of Computer Science and Engineering, Frankfurt University of Applied Sciences, 60318 Frankfurt am Main, Germany
[2]Escuela Superior de Ingeniería, Universidad de Cádiz, 11001 Cádiz, Spain
[3]School of Science and Technology, IE University, 28006 Madrid, Spain

Corresponding author: Fatima Sajid Butt (fatima.butt@fb2.fra-uas.de)

This work was supported in part by the PhD-research program of the Faculty of Computer Science and Engineering Fb2 of Frankfurt University of Applied Sciences. The research of DGU is supported in part by the Spanish Agencia Estatal de Investigaci'on under grants PID2021-122154NB-I00 and TED2021-129455B-I00, and by a 2021 BBVA Foundation project for research in Mathematics. He also acknowledges support from the EU under the 2014–2020 ERDF Operational Programme and the Department of Economy, Knowledge, Business and University of the Regional Government of Andalusia (project FEDER-UCA18-108393).

ABSTRACT Many recent studies have focused on the automatic classification of electrocardiogram (ECG) signals using deep learning (DL) methods. Most rely on existing complex DL methods, such as transfer learning or providing the models with carefully designed extracted features based on domain knowledge. A common assumption is that the deeper and more complex the DL model is, the better it learns. In this study, we propose two different DL models for automatic feature extraction from ECG signals for classification tasks: A CNN-LSTM hybrid model and an attention/transformer-based model with wavelet transform for the dimensional embedding. Both of the models extract the features from time series at the initial layers of the neural networks and can obtain performance at least equal to, if not greater than, many contemporary deep neural networks. To validate our hypothesis, we used three publicly available data-sets to evaluate the proposed models. Our model achieved a benchmark accuracy of 99.92% for fall detection and 99.93% for the PTB database for myocardial infarction versus normal heartbeat classification.

INDEX TERMS Electrocardiograph, benchmark testing, fall detection, time series analysis, machine learning, deep learning, LSTM, CNN, attention, transformer, PTB XL.

I. INTRODUCTION

According to [1], in the United States of America alone, the leading cause of death for men and women irrespective of the racial and ethnic groups is heart disease. Hence, a timely and accurate diagnosis of the heart conditions is of vital importance. An Electrocardiogram (ECG) is a well-grounded method used for measuring and evaluating the performance of the cardiovascular system. Several techniques exist in both literature and practice to evaluate the ECG signals in different manners. It is one of the most important parameters that indicate a person's physiological well-being and is extensively used to evaluate the cardiac situation of the patients. It has been widely used for different purposes such as to get an overview of the health of a human heart, for bio-metric purposes, and for fall detection and prevention as described in [2]. ECG is a non-invasive method for evaluating the health of the human cardiovascular system. It can detect many heart diseases such as atrial fibrillation, myocardial infarction, AV block, and ventricular tachycardia, etc. It provides an insight into the central nervous system, particularly the autonomic nervous system. Many of the automatic classification techniques using deep learning for ECG use either very deep neural networks or a pre-trained neural network that require either the weights set up to a configuration after being trained on an immense amount of similar data sets. Another approach is to pre-process the data sets by applying some filtration or feature extraction which is based on data domain knowledge and then fed into a neural network to train this. All of the above–mentioned steps involve an explicit understanding of the domain and the pre-process itself. [3] overviews

The associate editor coordinating the review of this manuscript and approving it for publication was Gerardo Flores.

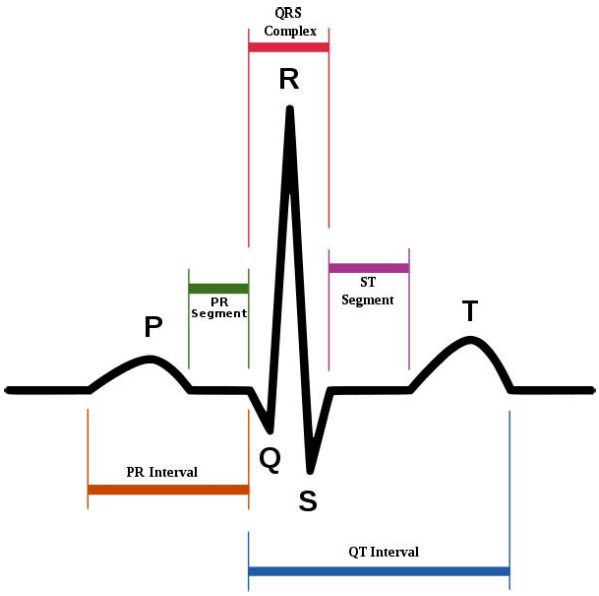

FIGURE 1. A Normal ECG.

the many ECG feature extraction techniques present in the literature.

Our work is motivated by the desire to design novel and simple models that avoid *any* feature selection and complex data pre-processing which necessitates domain knowledge, while on the other hand requiring *less* computation power but achieving at least state-of-the-art accuracy. In this paper we propose two architectures to achieve these goals including out-performance of benchmarks and analysis of the statistical evidence of our claims.

II. ELECTROCARDIOGRAM - A TIME SERIES

This section presents an introduction to ECG signals and the importance of using automatic ECG classification techniques. An ECG is a physiological signal that is measured as the potential difference between the electrodes placed on the body surface. The cardiac impulse passes through the heart causing the electrical current to spread from the heart to adjacent tissues. A small current extends to the surface of the body. The electrodes placed on the skin can effectively detect the current on opposite sides of the heart, and record the electrical potentials generated by the current. A normal ECG consists of five major deflections called P, Q, R, S, and T waves, which constitute a single cardiac rhythm as shown in Fig. 1. The P wave lasts about 0.08 s and is the smallest, followed by the large QRS complex which lasts between 0.08 s and 0.10 s. The end of the cardiac cycle is marked by a T wave that lasts approximately 0.16 s. A single waveform varies depending on the size of the heart and the conductive properties of the body which in turn gives the waveform a unique pattern per person [4]. ECG has not only been used to monitor and evaluate the cardiovascular system but has also been used as a biometric identifier [5], a predictor of gender and age as described in [6], and for detecting fall activities as in [7].

The disruption of blood flow to the muscle layer of the heart causes a cardiovascular condition called a myocardial infarction (MI). This disruption is mostly due to the build-up of the plaques in the arteries which result in reduced blood flow to that part of the heart muscle. MI is called a silent heart attack because the patient is not aware of the condition unless they suffer from a heart attack. An early diagnosis of MI is therefore of vital importance as it would help the patients to get timely treatment hence preventing the high percentage of mortality associated with it. Due to the small amplitude (millivolts), the manual interpretation of ECG signals is time-consuming and prone to errors. This limitation can be mitigated by an automatic diagnosis of heart conditions based on the signals. Our study aims to work towards automation of the cardiovascular disease diagnosis from ECG signals.

In this study, we propose two methods to automatically extract features from a time series and then feed those features into another deep learning model for classification. First, a hybrid model for multiple ECG classification tasks is proposed as an alternative to many complex models that require many pre-processing steps before the actual training. We experimented with a robust hybrid deep learning model for the ECG classification tasks, which proved to outperform many state-of-the-art complex models and achieve similar or even better accuracy with no pre-processing steps. The CNN placed in front of a LSTM also known as CNN-LSTM, has recently been used for multiple classification tasks; however, its use for ECG classification has not been systematically explored. The CNN model first searches for the features in high-dimensional input data and then after converting it into one-dimensional data, it is fed as an input to the LSTM model. The role of a CNN in this context is to act as an automatic feature extractor. Secondly, a novel attention/transformer model using wavelets for dimensional embedding is introduced to improve the efficiency of the classification process. As it has less trainable parameters than CNN-LSTM it has advantages in terms of (training) performance as shown in Table 8. As a bonus, we also evaluated both models also on the data for fall detection.

III. RELATED WORK AND OUR CONTRIBUTION

Several recent studies have focused on automatic ECG classification. Among the several different techniques present for ECG classification, deep learning has gained popularity in recent times. This is mainly owing to its automatic feature learning and the availability of large public data sets. Many deep learning techniques use feature extraction as an essential pre-processing step before feeding the data to the neural network. The most common feature extraction techniques for ECG classification are continuous wavelet transform (CWT), discrete cosine transform (DCT) [8], Pan-Tompkins algorithm [9] and discrete wavelet transform (DWT) [10]. One of the major disadvantages of using wavelet transform as a feature extractor is that the complexity of the process increases with the increase in decomposition level. All feature extraction processes require some domain knowledge in

order to efficiently extract relevant features from the data. Therefore, we aim to explore the research question of whether a similar state-of-the art result can be achieved with no pre-processing and with a simpler model architecture in an efficient manner in terms of resources and computation.

Our contributions to this study are twofold: First, we introduce a CNN-LSTM architecture that surpasses many complex and pre-trained models that have been optimized for single data sets on multiple data sets at the same time. Second, to further optimize the automatic feature extraction, we introduce a novel embedding technique for an attention/transformer encoder architecture that uses discrete wavelet transform to extract features from the ECG time series and feeds them to the attention mechanism. In addition, we provide statistical evidence for the significance of the performance figures reported by the two models proposed by us.

In the following sub-sections, we present the state of the art in the related work and highlight our contributions.

A. CNN-LSTM ARCHITECTURES

Jambukia et al. [4] presented an overview of the ECG classification of different types of arrhythmias. Another current review on deep learning methods for ECG arrhythmia classification [11] deduced that among the many deep learning models, CNNs and LSTMs were among the most effective for learning arrhythmia in ECG classification tasks. The use of the CNN-LSTM architecture for classification is not entirely novel. Socher et al. [12] proposed a model for 3d object classification that combines a CNN with an RNN. They concluded that the CNN provides the translation variance for lower-level features whereas RNNs can learn the interactions and compositional features in the data. Zheng et al [13], transformed the data acquired by a three-axis accelerometer into an image format and then used a CNN with three convolution layers to classify human activities. XIA et al. [14] used CNN after a LSTM layer to classify human activity recognition (HAR) with an accuracy of 95.85%. Ordóñez et al. [15] proposed an activity recognition classifier that combines a deep CNN and dense layers. In [16] the authors proposed a 1-D CNN for the classification of cardiac arrhythmia, and in [17], a 34-layer convolutional neural network is used for classification of cardiac arrhythmia exceeding the performance of board-certified cardiologists. However, few studies have focused on hybrid CNN-LSTM models for ECG classification. Studies like [18], [19], [20], and [21] have implemented CNNs and their variants for ECG classifications. [22] used RNNs to classify the ECG signals. The use of LSTM-based approaches is also beneficial for other cardiac signal analyses. Reference [23] construct a bidirectional LSTM for the analysis of blood flow dynamics from static CT angiographic images. In [24] a restricted Boltzmann machine and deep belief networks were used for detection of ventricular and supraventricular heartbeats using single-lead ECGs. For a general overview of deep-learning techniques in cardiovascular image analysis, see the survey [25].

In our study, we not only performed multiple classifications with CNN-LSTM model for ECG but also worked with three different ECG data sets including data for fall detection to present a proof of concept that CNN placed in front of LSTM surpasses many complex and pre-trained models.

B. ATTENTION AND TRANSFORMER ARCHITECTURES

The seminal paper by Vaswani et al. ''Attention is All you Need'' [26] has triggered an enormous number of successful applications of attention mechanisms and transformer architectures in deep learning.

The main idea behind attention-based transformer architectures is to replace the recurrence mechanisms used in LSTMs and the convolutions used in convolution networks to extract features entirely using an alternative so-called self-attention mechanism. This mechanism is shown in eq. (1) and computes the correlation between the input values among each other and can be interpreted as an associative memory using ideas from statistical physics, see [27]. Replacing the (serial) recurrence mechanism with the standard matrix algebra of the (self-) attention mechanism has a number of advantages for parallelization capabilities and the performance of classification tasks.

However, the vast majority of research has been and still is focused on the natural language processing (NLP) domain. Little research has been carried out on the application of attention-based architectures in other domains, such as time series analysis. One of the first papers in this regard is LSTNet by Lai et al. [28], where the authors introduced long- and short-term time-series networks (LSTNet) using the convolutional neural networks and recurrent neural networks to extract short-term local dependency patterns and to discover long-term patterns for time series trends. Shih et al. [29] applied an attention mechanism to multivariate time series data in three medical domains. Song et al. [30] have applied attention models to clinical time series analysis. A systematic and comprehensive analysis and study of utilizing attention mechanisms, however, in the time-series domain is still required. The application of transformers in the domain of ECG classification can be found in [31].

One of the shortcomings of the self-attention mechanism preventing its application for e.g., time-series is the $\mathcal{O}(n^2)$ complexity with regards to the length of the input vector, i.e., the length of the time series in our case. To address this problem LinFormer has been introduced by Wang et al. in [32]. Linformer is the first theoretically proven linear-time transformer architecture and henceforth might be suitable also for long time series. The linear scaling is achieved by discovering that self-attention is low rank, and henceforth projecting information on a low rank constant sub-dimensions achieves to decouple from the $\mathcal{O}(n^2)$ scaling. Recently Rabe and Staats [33] have proposed an algorithmic solution to at least reduce the memory (but not the time) complexity from $\mathcal{O}(n^2)$ to $\mathcal{O}(n)$.

In this paper, we propose a novel attention architecture using projection on discrete wavelet components as a

means to address the $\mathcal{O}(n^2)$ problem and for dimensional embedding. Moreover, the results show that using this technique, *attention-only* architecture is on par with or even outperforms more complex models and has several additional advantages such as e.g., better run-time performance.

IV. ALGORITHMS

This section provides a brief overview of the algorithms and the technologies that were used during the course of this study and also presents the state of the art in the respective technologies.

A. CNN-LSTM MODEL

We define some basic terms related to the convolutional neural network and LSTM for clarity in the following section.

1) CNNs AND LSTMs

Convolutional neural networks, introduced as LeNet in 1989 by LeCunn, have revolutionized the field of image recognition and are among the most prominently used deep neural networks. They were named after the linear matrix operation called convolution. Since convolution is a linear operation, the convolution layer is often followed by a non-linear layer. Although introduced earlier, it gained popularity after its application as the first deep neural network applied for object recognition in the ImageNet Large Scale Visual Recognition Competition (ILSVRC) in 2012. AlexNet was proven to excel on the largest computer vision data set as compared to contemporary methods. Recently, [34] presented a state-of-the-art review of the recent deep CNNs architectures. The individual CNN components were explained in [35] in a structured way. The most common architectures of CNNs include an input layer, a convolution layer followed by a pooling layer, a drop-out layer, and a fully connected layer followed by an output layer. The number of layers and their layout can change depending on different problem sets. The convolution operation[1] itself is given by:

$$V_{i,j} = X * W_{i,j} + b = \sum_{L} X^L * W^L_{i,j} + b$$
$$= \sum_{L} \sum_{k,l} X^L_{kl} * W^L_{i+k,j+l} + b$$

where X or X^L resp. denote the L-th input matrix. W is the convolution kernel matrix, b is the bias, and $V_{i,j}$ is the output matrix after convolution.

CNN's are known for their excellent feature extraction capability. One of the most salient features of CNN is its translation invariance. Therefore, it can extract features irrespective of the spatial context. Though it has proven to be beneficial in image recognition, its application and usefulness in time series are yet to be fully exploited. Cases, where the historical context is relevant for classification, would not work well with CNN alone, as it does not carry any information about the history of the time series. The CNNs initially extract the local features in the sub-regions of the time series and then the information is merged in later stages to detect the higher order features. We applied 1D convolution to the time series using both univariate and multivariate data sets. The ECG Human activity recognition (HAR) data set and PTB diagnostic data set contained one feature each, so the 2D convolutional operation would not be suitable as it will incorrectly convolve across multiple time series. Long short-term memory (LSTM) networks–a variation of recurrent neural networks (RNNs)–were introduced by Hochreiter [36] in 1997. They tend to present a solution to the common problem associated with RNNs called vanishing and exploding gradients. In principle, classical RNNs can keep track of long-term dependencies in the sequences. However, in practice, during the backpropagation phase of training, these long-term gradients either vanish or explode owing to the successive multiplicative operations. An LSTM consists of a chained loop structure. Each LSTM unit is made up of an input gate, an output gate and a forget gate. The LSTMs keep the long-term memory by maintaining a cell state that sustains a part of the information from earlier states by forgetting and/or applying increment operations on the previous states. Adding a CNN in front of an LSTM helps to feed the LSTM the features from CNN which were extracted from the time series.

[1] in the two dimensional case–the one-dimensional case is analogous.

2) CNN-LSTM ARCHITECTURE AND ALGORITHM

Fig. 2 and Fig. 3 provide a more graphical overview of our model. Initially (1*N) time series with N time stamps are convolved with k filters each of size M*1. Subsequently, the k feature maps each of size (N-M)+1 time stamps are generated which are passed through a dropout layer followed by the max pooling layer and later fed into the LSTM layer where the encoded extracted features are fed into it from the CNN. The LSTM unit is followed by a fully connected or dense layer that applies softmax as an output function to classify the input time series into one of the output classes.

Algorithm 1 Classification of ECG Signals With Raw Signals Using CNN-LSTM

Input: A time series ECG raw data *ts*
Output: The classified label *l*
1: *ts* ← RAW_VALUE_EXTRACTION(*ts*)
2: *features* ← CNN(*ts*)
3: *l* ← LSTM_CLASSIFICATION(*features*)
4: **return** *l*

Algorithm 1 outlines the algorithms for extracting features from the ECG signal and classifying them using a CNN-LSTM model. The number of CNNs and LSTMs can be varied but we used a maximum of five 1-d convolution layers in front of the three LSTM layers.

```
==========================================================
Layer (type:depth-idx)                   Param #
==========================================================
CNN_LSTM                                 --
├─CNN: 1-1                               --
│    └─ReLU: 2-1                         --
│    └─Sequential: 2-2                   --
│    │    └─Conv1d: 3-1                  561
│    │    └─BatchNorm1d: 3-2             374
│    │    └─ReLU: 3-3                    --
│    └─Sequential: 2-3                   --
│    │    └─Conv1d: 3-4                  24,000
│    │    └─BatchNorm1d: 3-5             128
│    │    └─ReLU: 3-6                    --
│    │    └─MaxPool1d: 3-7               --
│    └─Sequential: 2-4                   --
│    │    └─Conv1d: 3-8                  8,256
│    │    └─BatchNorm1d: 3-9             128
│    │    └─ReLU: 3-10                   --
│    │    └─MaxPool1d: 3-11              --
│    └─Sequential: 2-5                   --
│    │    └─Conv1d: 3-12                 8,256
│    │    └─BatchNorm1d: 3-13            128
│    │    └─ReLU: 3-14                   --
│    │    └─MaxPool1d: 3-15              --
│    └─Sequential: 2-6                   --
│    │    └─Conv1d: 3-16                 8,256
│    │    └─BatchNorm1d: 3-17            128
│    │    └─ReLU: 3-18                   --
│    │    └─MaxPool1d: 3-19              --
│    └─Sequential: 2-7                   --
│    │    └─Conv1d: 3-20                 8,256
│    │    └─BatchNorm1d: 3-21            128
│    │    └─ReLU: 3-22                   --
│    │    └─MaxPool1d: 3-23              --
│    └─Sequential: 2-8                   --
│    │    └─Conv1d: 3-24                 16,512
│    │    └─BatchNorm1d: 3-25            256
│    │    └─ReLU: 3-26                   --
│    │    └─Dropout: 3-27                --
│    │    └─MaxPool1d: 3-28              --
├─LSTM: 1-2                              264,192
├─Linear: 1-3                            2,562
==========================================================
Total params: 342,121
Trainable params: 342,121
Non-trainable params: 0
==========================================================
```

FIGURE 2. CNN-LSTM Model for PTB DB.

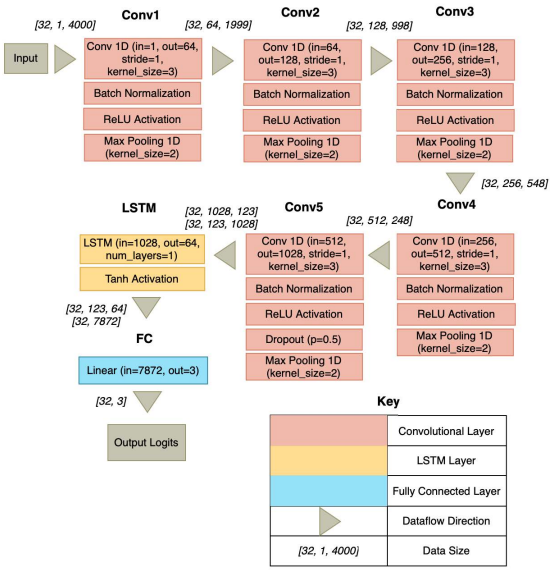

FIGURE 3. Final CNN-LSTM Architecture for Fall and HAR using ECG signals.

B. ATTENTION MODEL

For reader's convenience, we recall the basic definitions of the attention mechanism following [26] and the notation therein.

1) ATTENTION

Attention is defined as

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{1}$$

where Q, K and $V \in \mathbb{R}^{n\times d_k}$ are input embedding matrices, n is the length of the (time) series, and d_k is the embedding dimension, resp.

The transformer uses Multi-Head Self-Attention (MHA) allowing the model to jointly attend to information at different positions of the time-series or different semantics of the domain. MHA is defined as

$$\text{MultiHead}(Q, K, V) \tag{2}$$

$$= \text{Concat}\left(\text{head}_1, \text{head}_2, \ldots, \text{head}_h\right) W^O, \tag{3}$$

where h is the number of heads. Each head is defined as

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{4}$$

$$= \text{softmax}\left[\frac{QW_i^Q(KW_i^K)^T}{\sqrt{d_k}}\right]VW_i^V, \tag{5}$$

where W_i^Q, $W_i^K \in \mathbb{R}^{d_m\times d_k}$, $W_i^V \in \mathbb{R}^{d_m\times d_v}$, and $W^O \in \mathbb{R}^{hd_v\times d_m}$ (projection onto the output) are learned matrices and d_k, d_v are hidden dimensions of projection subspaces. For simplicity in the sequel, we drop the differentiation between d_m, d_k and d_v and refer to them by d.

The matrices Q, K and V are usually referred to as query, key and value matrices to remind of the associative memory architecture of a transform, compare e.g., also the analysis in [27].

2) ATTENTION AND DIMENSIONAL EMBEDDING

For applying the attention mechanism to time-series one has to decide on the proper dimensional embedding, i.e., on the dimension of the embedding subspace and on the embedding transformation. We recall that in the domain of ECG the "natural" dimension is small. For instance, the signals are one-dimensional if a one-dimensional channel (single lead) is used (as is the case in this paper for the attention/transformer model, i.e., Algorithm 2). Even if multi-channel ECGs are used usually the number of channels is limited to a small number of 3 to maximally 12 channels. Henceforth, if we used the channel as the embedding, the dimension would be 1 in our case, i.e., $d = 1$. This is way too small to capture interesting patterns and, indeed, a test showed that the gradient descent does not converge, but stays constant after one or two initial updates. Furthermore, as depicted in the previous section, the self-attention suffers from an $\mathcal{O}(n^2)$ problem. We propose the following architecture to solve both problems simultaneously:

1) Assuming $m \ll n$. For simplicity of the notation, we assume without loss of generality that n is divisible by m, i.e., $n = mw$. This effectively segments the time-series n into n_m "windowed" segments of length w, where $m \in 1, \ldots k$ with $k := n/w$. If n is not divisible

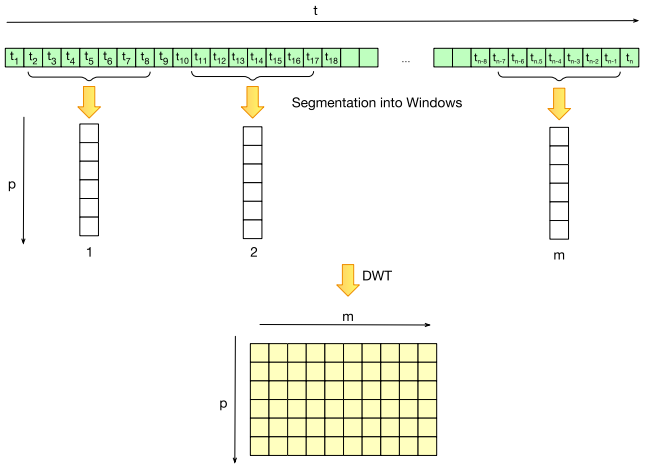

FIGURE 4. Dimensional Embedding.

without remainder, we could fill the time series with zeros (padding).

2) For each "windowed" sub- time-series t_{n_k} we calculate the decomposition to a chosen (fixed) wavelet by performing a discrete wavelet transformation (DWT), see below. Assuming that the result of applying the DWT is in dimension p, we have transformed the input from $\mathbb{R}^n$ into $\mathbb{R}^{m\times p}$ depicted in Fig. 4.

Remark 1: In this paper, we propose a deterministic embedding using DWT rather than a learned, randomly initialized embedding, which is an alternative approach that has been used in other attention architectures in the past. This should–in theory–require less training data–a conjecture that we want to validate in future work using synthesized data.

Remark 2: Note, that within the context of dictionary-based learning, a deterministic embedding using DWT could be considered as a "predefined analytical dictionary" [37]. Contrary to a fixed feature design using wavelet components, however, an embedding with an attention/transformer architecture has a flavor of learning the representation dynamically from the data as the proper amount of *attention* is learned from the data indeed. A systematic investigation of these aspects is deferred to future work, too.

3) DISCRETE WAVELET TRANSFORMATION (DWT) FOR DIMENSIONAL EMBEDDING

For the reader's convenience, we recall a few well-known definitions and theorems wavelet theory following [38] and [39] and using the notation from [40]. We consider signals as real-valued functions. We call a function $\psi \in L^2(\mathbb{R})$ an orthonormal wavelet if dyadic translations and dilations of $\psi_{jk}(x) = 2^{\frac{j}{2}}\psi\left(2^j x - k\right)$ constitute a Hilbert space and if in addition, it satisfies a regularity (admissibility) condition, namely $\int_0^\infty |\psi(t)|^2 \frac{dt}{t} < \infty$, ensuring convergence and $\int \psi(t)\,dt = 0$. The function ψ is usually called the *mother wavelet* and its child wavelets are defined as $\psi_{j,k}(t) = \frac{1}{\sqrt{2^j}}\psi\left(\frac{t-k2^j}{2^j}\right)$. A projection of a function $x(t)$ onto $\psi_{j,k}$ is

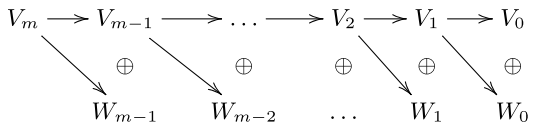

FIGURE 5. Wavelet Transform Pyramid.

then given by

$$\gamma_{jk} = \int_{-\infty}^{\infty} x(t)\frac{1}{\sqrt{2^j}}\psi\left(\frac{t-k2^j}{2^j}\right)dt. \tag{6}$$

The crucial idea is to decompose the space $L^2(\mathbb{R})$ into resolution spaces of different resolutions. First, the *resolution space* V_0 is defined as the space of piece wise constant functions on subintervals of $[n, n+1]$ with $n = 0, \ldots, N$. If we define the corresponding step function

$$\phi(t) = \begin{cases} 1, & if \quad 0 \le t < 1 \\ 0 & otherwise, \end{cases} \tag{7}$$

then V_0 has dimension N, and the N functions $\boldsymbol{\phi}_0 := \{\phi(t-j)\}_{j=0,\ldots,N-1}$ constitute an orthogonal basis. Analogously the *refined resolution spaces* V_k are defined as the spaces of functions constant on each sub-interval $[n/2^k, (n+1)/2^k]$. This yields a nested sequence of embedded spaces

$$V_0 \subset V_1 \subset V_2 \subset \ldots \subset V_k. \tag{8}$$

We denote the orthogonal complements of V_{k-1} in V_k as W_{k-1}, i.e., $V_k = V_{k-1} \oplus W_{k-1}$ and call it the *detail space*. This yields an orthogonal decomposition at level k as follows:

$$V_k = V_0 \oplus W_0 \oplus W_1 \oplus \ldots \oplus W_{k-1} \tag{9}$$

Then the k-level *discrete wavelet transformation* (DWT) is defined as the change of coordinates from $\boldsymbol{\phi}_k$ to $(\boldsymbol{\phi}_0, \boldsymbol{\psi}_0, \boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_{k-1})$, where $\boldsymbol{\phi}_k := \{\phi_{jk}\}_{j=0,\ldots,N-1}$ and $\boldsymbol{\psi}_k := \{\psi_{jk}\}_{j=0,\ldots,N-1}$, resp. denote the family of functions obtained from the mother wavelet.

This yields a filter bank interpretation of DWT, wherein each step the signal is decomposed into an averaged and a detailed signal using low-pass and a high-pass filter depicted graphically in Fig. 5.

Thus, any signal can be decomposed into an averaged and a detailed signal, namely V_0 and W_0. Due to the recursive nature, this can be extended to any desired level k. Please note, that due to the dyadic nature the data size is reduced by a factor of 2 in each step. For further details, we refer to [38] and [39] as well as e.g. [40].

We used Haar and Daubechies wavelets as well as symlets (symmetrized version of Daubechies wavelets) as mother wavelet.

4) TRANSFORMER ARCHITECTURE AND ALGORITHM

The DWT can be used not only for dimensional embedding but also, for noise reduction as one can ignore some or all

TABLE 1. Attention Models.

	Wavelet	Spaces	Dim Embedding	Hid Dim	Frequency	Epochs	Accuracy
ATT2	db8	V_0	13	150	0.1k	1k	98.14
ATT3	db8	V_0	12	150	0.1k	1k	98.83
ATT4	db8	$V_0 \oplus W_0$	24	150	0.1k	1k	99.18
ATT5	db8	$V_0 \oplus W_0$	24	150	0.1k	1k	99.11
ATT6	db8	$V_0 \oplus W_0$	26	250	0.1k	1k	99.45
ATT7	sym6	$V_0 \oplus W_0$	22	250	0.1k	1k	99.38
ATT8	db8	$V_0 \oplus W_0$	26	250	10k	1k	99.59
ATT9	db8	$V_0 \oplus W_0$	26	250	10k	2k	99.59
ATT10	db5	$V_0 \oplus W_0$	20	250	10k	1k	99.24
ATT11	db5	$V_0 \oplus W_0 \oplus W_1$	28	250	10k	1k	99.73

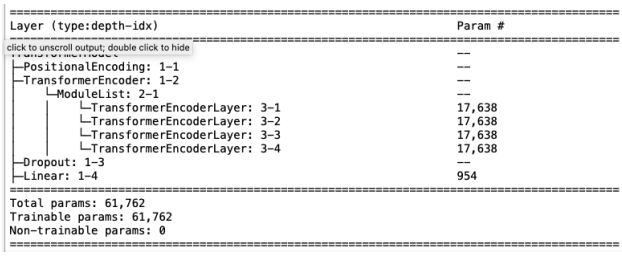

FIGURE 6. Transformer Architecture of Model att11.

coefficients for the detailed spaces. To explore the impact, we tried several configurations, see Table 1.[2]

We deployed a transformer architecture with one head, four transformer encoder layers, and a dimension of 150 or 250 hidden units of the feed-forward network. The network's architecture of the best model, att11, is illustrated in Fig. 6.

The original input tensor has a dimension of [14552, 1, 187] corresponding to 14522 data rows, 1 feature, and a sequence length of 187. The sequence was split into 17 chunks, with a window length of 11. Each subsequence of length 11 was converted using the DWT. For instance, for db8 for att7, this leads to a transformed tensor of [14552, 22, 17]. (Note, that due to boundary effects, the embedding dimension is not always a multiple of 11.).

As a positional encoding, we tried the usual Fourier encoding and used frequencies f of $f = 100$ and $f = 10.000$, resp. While adding positional encoding is questionable after embedding using DWT, we experimentally found the results to be improved by a small amount.

All models were trained with a batch size of 256 and a learning rate of 0.001 using the Adam optimizer [41]. Algorithm 2[3] layouts the algorithms for extracting features from ECG data and classifying them using a transformer/attention model.

[2]Frequency refers to the frequency of the positional encoding. It should be remarked, that model att5 differs from att4 by an additional residual connection.

[3]Note, that for simplicity of notation we use the expression $(V_0 \oplus \ldots \oplus W_m)$ from the decomposition in equation 9 generically, i.e., some of the components of $(V_0 \oplus \ldots \oplus W_m)$ might be empty.

Algorithm 2 Classification of ECG Signals With Raw Signals Using Attention/Transformer

Input: A time series ECG raw data ts
Output: The classified label l

$X \leftarrow$ RAW_VALUE_EXTRACTION(ts)
$(V_0 \oplus W_0 \oplus W_1 \oplus \ldots \oplus W_m) \leftarrow$ DWT(X)
$(V_0 \oplus \ldots \oplus W_m) \leftarrow (V_0 \oplus \ldots \oplus W_m) + POS\_ENC(V_0 \oplus \ldots \oplus W_m)$
for $i \in$ Layers **do**
  $X \leftarrow$ TRANSFORMER$_i(V_0 \oplus \ldots \oplus W_m)$
end for
$l \leftarrow$ LINEAR_FEED_FORWARD(X)
return l

TABLE 2. Generic Runtime Complexity Analysis.

Layer Type	Layer Compl.	Seq. Ops	Max Path Length
Attention	$\mathcal{O}(n^2d + nd^2)$	$\mathcal{O}(1)$	$\mathcal{O}(1)$
RNN/LSTM	$\mathcal{O}(nd^2)$	$\mathcal{O}(n)$	$\mathcal{O}(n)$
CNN	$\mathcal{O}(knd^2)$	$\mathcal{O}(1)$	$\mathcal{O}(\log_k(n))$

TABLE 3. Runtime Complexity Analysis for Algorithms 1 and 2.

Algorithm	Total Complexity	Max Path Length
Alg 1	$\mathcal{O}(n^2d^2)$	$\mathcal{O}(n + \log_k(n)) = \mathcal{O}(n)$
Alg 2	$\mathcal{O}(n^2d + nd^2)$	$\mathcal{O}(1)$

C. COMPLEXITY ANALYSIS

1) RUNTIME COMPLEXITY ANALYSIS

In general, the runtime complexity for attention based RNN/LSTM and CNN architectures are known as depicted in Table 2, see e.g. [26],[4] where n denotes the sequence length, d the embedding dimension, and k the size of the kernel (in case of CNN).

Note, that the maximum path length measures the maximum length between any two input and output positions in the networks. Shorter path length makes it easier to learn long-range dependencies.

Considering, that the dimensional embedding using wavelets is of $\mathcal{O}(n)$ and has to be computed only once and henceforth can be ignored compared to $\mathcal{O}(n^2d + nd^2)$, we conclude from Table 2 the complexity of our algorithms as depicted in Table 3.

From this analysis, we can conclude that Alg 1 is always inferior to Alg 2 in terms of algorithmic complexity. In addition, the transformer can be easily paralleled (typically on a GPU), contrary to a CNN-LSTM.

Please note, that the above analysis assumes that the matrix multiplication of two matrices $\mathbf{A} \in \mathbb{R}^{nm}$ and $\mathbf{B} \in \mathbb{R}^{ml}$ is in $\mathcal{O}(nml)$, which corresponds to a naive implementation of matrix multiplication. Although this can

[4]Note, that we have added the complexity caused by the query matrices which were omitted in [26].

TABLE 4. Total Number of Samples in the ECG HAR Data Set [7].

Label	Total Count
Fall	500
Rest	474
DA	296
Total	1270

be improved, compare Strassen's algorithm [42], and–more recently–Josh Alman and Virginia Vassilevska Williams algorithm [43], these algorithms are normally *not* implemented in the machine learning frameworks used and henceforth the naive implementation as the usual convention is assumed.

2) MEMORY COMPLEXITY ANALYSIS

As for backpropagation, the weights optimized have to be kept in memory, to optimize the algorithms efficiently, we have essentially the same space as time complexity, particularly space complexity of any attention-based model of $\mathcal{O}(n^2)$.

V. DATA PREPARATION AND EXPERIMENTAL SETUP

Since our study includes multiple data sets, therefore this section explains the preparation steps taken for each data set and the overall experimental setup. Experiments were performed using a GPU server. All the experiments were implemented using the PyTorch library because of its supportive architecture with GPUs. The main aim of the experiments was fall and MI detection using ECG signals in an automated and efficient manner.

A. ECG DATA SET FOR FALL DETECTION

To the best of our knowledge, the ECG HAR data set is the only one for the detection of different human activities including falls, using ECG signals. It was originally collected by [7], as an experiment that was part of the study by [44]. It originally consisted of two classes: one for the ECG of a person falling from the bed and another one for the ECG of a resting person. It was later augmented with two more data sets, [45] and [46], by up-sampling the original data set. In addition to that, another augmentation method called slicing was applied to the data set. Slicing has been explained in detail in [47]. After the addition of new data sets, the final version has three classes namely: fall, rest, and daily activities.

The overview of the final class distribution in data set is depicted in Table 4.

In the previous experiments, the data set was filtered, converted to wavelet transform, and later to 3-D images called scalograms. These scalograms were first used to fine-tune and then train, a pre-trained AlexNet and GoogLeNet. The state-of-the art validation accuracy obtained for classification with this data set is 98.44%. This accuracy was obtained after applying extensive pre-processing to the data set. Our current model outperforms the state-of-the-art validation accuracy and achieves a 99.21% accuracy with no pre-processing and only fine tuning the ensemble model.

B. PTB DIAGNOSTICS DATA SET

After working with the ECG for falls and daily activities, the model had to be tested on a standardized data set that is publicly available. In the second set of experiments, a publicly available data set called the PTB diagnostic was used, which is freely available but is used as a standard for ECG classification tasks.

The original PTB data set consists of 549 records from 290 subjects which were aged 17 to 87 years, with a mean age of 57.2. A total of 209 subjects were males with a mean age of 55.5 and 81 females with a mean age of 61.6 (for 1 female and 14 male subjects; age was not recorded). Each record has 15 measured signals: the conventional 12 leads (i, ii, iii, avr, avl, avf, v1, v2, v3, v4, v5 and v6) together with the 3 Frank lead ECGs (vx, vy and vz). The data from lead II were used to train the model which outperforms the databases which even use all 12 lead data [48]. ECG beats were extracted using the method described in [18]. The data set used was divided into two classes: normal and abnormal (myocardial infarction). In the previous prominent study [49], all 12 leads were separately evaluated to determine which leads contributed the most to the classification. We used lead II of the data set to differentiate between healthy controls and that with myocardial infarction. Since only one lead of ECG was used in the previous two experimental phases, we used another publicly available data set and used all 12 of its leads to reaffirm the usefulness of the ensemble model for both uni-variate and multivariate data sets.

C. PTB XL DIAGNOSTICS

PTB-XL is one of the largest freely accessible ECG data sets available. It was collected over a span of seven years between 1989-1996. It was made publicly available in 2020 in a structured database by Physikalisch-Technische Bundesanstalt (PTB). The data set consists of a total of 21837 records of 12-lead ECG each comprising of 10 s. It is a gender-balanced data set with 52% male and 48% female records and an age range of 0-95 years. The data set consisted of various diagnoses and a large number of healthy controls as well [50]. PTB XL has a standardized set of pre-processing instructions for the data set. Because different labels are heavily imbalanced and imbalanced classes can introduce bias in the trained model, it is important to divide the data set in a way that each of the classes is represented equally in each subset. Stratified sampling was used to divide it into training-validation-testing data sets. The data set has multiple classification categories as shown in Fig. 7. The goal is to classify MI from other heart conditions, and models were trained for diagnostic superclass and myocardial infarction detection using Algorithm 1.

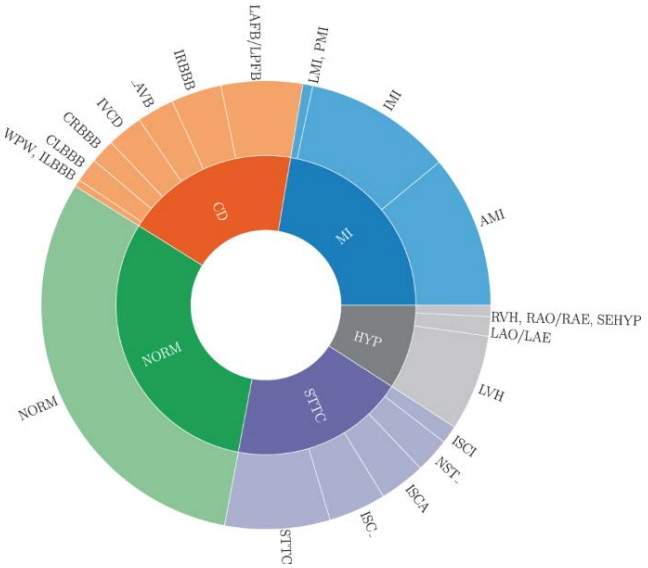

FIGURE 7. Class Distribution for PTB-XL Data Set.

VI. RESULTS

Several methods to evaluate a DL classifier exist in the literature. We evaluated our classifiers using the accuracy, area under the curve (AUC), confusion matrix, sensitivity, and specificity. The details of the results obtained by applying each of the algorithms and the respective data set is explained in this section. In the sequel, we present the results acquired for each data set and also compare our results to the other state-of-the-art work. An overview of the section is presented in Table 14.

A. ECG HAR DATA SET

The data set was initially trained using a plain LSTM to compare the model performance with the previous experiments. The data were fed into the model without any pre-processing. The LSTM initially yielded an accuracy of 49.80% which was increased to 54% by fine-tuning the hyperparameters. In the previous experiments, extensive pre-processing was carried out to extract the related features and then those features were fed into the model. Although that approach yields excellent accuracy, it is not automated. LSTMs have been shown to have a sense of previous timestamps or history in the time series, but CNNs have a superior feature extraction capability. To test our hypothesis, a CNN was placed on top of the LSTM layer. The accuracy immediately improved to 93%. After some fine tuning the hyperparameters and adjusting the number of CNN layers, the validation accuracy got better than the state-of-the-art results. A testing accuracy of 99.21%–100% was achieved and a validation accuracy of 99.21% was achieved. For the first data set, the results were almost perfect with a validation accuracy of 99.21% and a testing accuracy of 99.21%–100%. The previous work achieved similar accuracy but with transfer learning and pre-processing the signals by converting them into wavelet transforms and then into scalograms. This model achieves similar accuracy even by avoiding all those steps. The following Table 5 depicts the confusion matrix for the testing data set showing an almost perfect accuracy of 99.22%.

TABLE 5. Confusion Matrix for Fall Detection ECG Data Set using CNN-LSTM (Algorithm 1).

		Actual		
		DA	Fall	Rest
Predicted	DA	30	0	0
	Fall	1	52	0
	Rest	0	0	45

TABLE 6. Confusion Matrix for Fall Detection ECG Data Set using Attention (Algorithm 2).

		Actual		
		DA	Fall	Rest
Predicted	DA	29	1	0
	Fall	2	49	2
	Rest	0	0	44

Fall detection using ECG signals was also performed by applying Algorithm 2 to the HAR data set. Each sequence in the data set consists of 4000 time stamps. The initial tensor size was [1273, 1, 4000], which is in the format [total Sequences, number of features, sequence length]. Each of the ECG sequences was divided into 100 chunks of 40 time stamps each, and then the wavelet transform was calculated for each chunk resulting in a final dimension of [1273, 108, 40]. The model was trained in 403.39 s. This result was again achieved without any manual feature extraction or transfer learning model. The following Table 6 depicts the confusion matrix for the testing data set showing also the accuracy of 95.31%

B. PTB DIAGNOSTICS

Algorithm 1, i.e., CNN-LSTM was used to model the PTB diagnostic to differentiate normal from abnormal heartbeats. Previous studies have emphasized feature extraction before feeding into the neural network, or transfer learning where the model is initially trained with an existing data set and later on trained with the same learned weights on the desired data set such as in [51]. In the current benchmark for MI classification using PTB diagnostic, ConvNetQuake neural network model was adapted to achieve an accuracy of 99.44%. Similarly, heavy pre-processing, such as wavelet transformation [52], data balancing [53], and transfer learning [18], are used in the literature to achieve higher accuracy for ECG signal classification. In our study, no pre-processing of the individual readings was applied, and the model achieving 99.66% accuracy, exceeded the state-of-the-art accuracy for normal versus abnormal classification, which was previously 99.43%.

Algorithm 2, i.e., the attention /transformer model was also used to model the PTB diagnostic to differentiate between normal and abnormal heartbeats. This yielded an accuracy of

TABLE 7. Confusion Matrices for PTB Data Set.

		Predicted			
		Algorithm 1		Algorithm 2	
		NORMAL	ABNORMAL	NORMAL	ABNORMAL
Actual	NORMAL	371	1	372	1
	ABNORMAL	2	1081	3	1080

99.73%, a precision of 99.73%, a sensitivity of 99.2%, and a specificity of 99.91%.

The confusion matrices of both algorithms are depicted in Table 7 below.

In comparison to CNN-LSTM, i.e., Algorithm 1, the attention model with wavelet embedding has been shown to be more efficient as it uses fewer parameters and less training time as compared to the CNN-LSTM model as shown in Table 8. However, the time and number of parameters for Algorithm 2 might increase eventually with the increase in the number of attention heads and encoder layers. A comparison between the reference parameters used across the state-of-the-art similar work and our work is shown in Table 8. However, it must be noted that Table 8 is not complete because not all parameters can be found for all related work and NA in the table refers to not available.

Multiple metrics were used to evaluate the model performance. The terms *tp*, *fp*, *tn*, and *fn* refer to the true positives, false positives, true negatives, and false negatives respectively. In medical terminology, true positive would refer to the medical condition being diagnosed, so *tp* in our context refers to the diagnosis of MI. The performance metrics were calculated using the following formulas:

$$\begin{aligned} Accuracy &= (tp + tn)/(tp + fp + fn + tn) \\ Precision &= tp/(tp + fp) \\ Sensitivity &= tp/(tp + fn) \\ Specificity &= tn/(tn + fp) \end{aligned}$$

The results for our leading models are summarized in Table 9.

Fig. 8 and Fig. 9 show examples of the training accuracy and losses for the PTB data set, respectively.

C. PTB XL DIAGNOSTICS

Since PTB XL is a relatively new data set, many recent studies using this data set have adapted it for different classification tasks such as super diagnostic, sub-diagnostic, and form etc. In our study, five super diagnostic (SD) classes were classified. A validation accuracy of 75.70% and a testing accuracy of 74.33% were achieved. An AUC score of 0.8395 was obtained, see Table 10.

A direct comparison to the state-of-the-art is not very straightforward for PTB-XL mainly because it is a newer

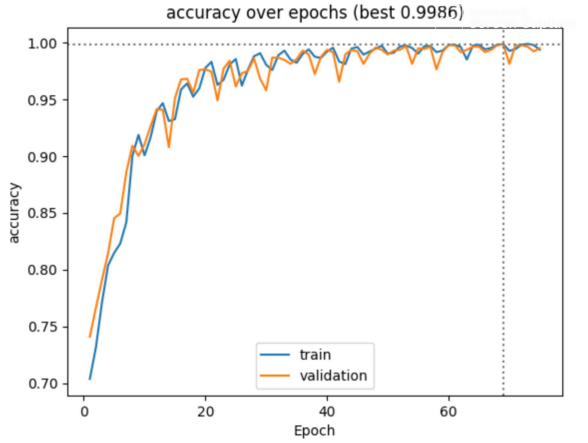

FIGURE 8. Training and Validation Graph over Epochs for the PTB Data Set for Algorithm 1.

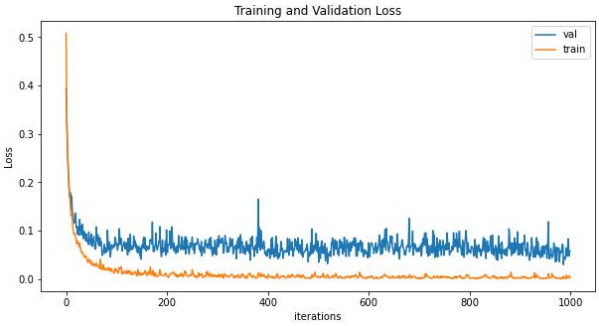

FIGURE 9. Training and Validation Loss for the PTB Data Set for Algorithm 2.

data set and many new studies that use it focus on different classification tasks. Although our results for this data set do not exceed the state-of-the-art accuracy, they are still comparable. Virginia et al. [54] and Martin et.al [55] achieved accuracies of 90.8% and 77.12%, respectively, for MI classification respectively. Similarly, Śmigielet et al. [56] achieved an accuracy of 78.0–75.2% for five-class classification.

Our Algorithm 1 applied for myocardial infarction detection using the PTB XL data set, yields a validation accuracy of 90.94% when it is MI versus the normal class in super diagnostic, and 91.27% when it is MI vs other four superclasses. Please note the AUC as a metric was calculated only for PTB-XL data set as it is widely used for comparison of this particular data set in the literature. The experiment with Algorithm 2 on multi-lead data sets such as PTB XL is still in a preliminary stage and requires further investigation on how to merge the natural domain dimension of multi-lead with the dimensional embedding technique. However, similar results for another research project of ours yield promising results for the Algorithm 2 with a multidimensional or multi-featured data set, see [57]. Henceforth these experiments have not been mentioned and as work in progress will be published in a future contribution.

TABLE 8. Parameter Comparison between State of the Art and Our Work.

Work	Total Parameters	Learning Rate	Time to Train	Training Hardware	Batch Size	Epochs	Optimizer
Liu et al. [58]	NA	0.001 - 0.000001	3557.26 s	2 NVIDIA Titan Xp GPUs	32	100	NA
Feng et al. [59]	NA	0.1 - 0.0001	1500s	Intel Core i5-4590@3.30GHz	32	100	Adam, rmsProp, SGD
Wang et al. [20]	NA	0.001	NA	2 NVIDIA 2080Ti GPUs	12	10	NA
Rai et al. [53]	NA	0.001	1263 s - 2285 s	i5 core, NVIDIA graphics card	128	100	Adam
Our work (HAR): Alg. 2	**461,583**	0.00002	**403.39**	NVIDIA A100-PCIE-40GB	20	500	Adam
Our work (HAR): Alg. 1	2,471,503	0.001	80.56s	NVIDIA A100-PCIE-40GB	64	100	Adam
Our work (PTB): Alg. 2	**61762**	0.001	**823.54s**	NVIDIA A100-PCIE-40GB	256	1000	Adam
Our work (PTB): Alg. 1	210,025	0.001	1923.87s	NVIDIA A100-PCIE-40GB	64	1000	Adam

TABLE 9. Metrics for Leading CNN-LSTM and Attention.

		Predicted	
		CNN-LSTM	ATTENTION
Actual	Accuracy	99.79	99.73
	Precision	99.91	99.91
	Sensitivity	99.81	99.72
	Specificity	99.73	99.73

TABLE 10. Performance Indicators.

Classification classes	Val. Accuracy	Testing Accuracy	AUC score
Super-diagnostic Classes	75.70%	74.33%	0.8395
MI vs. Normal Class	90.94%	89.1%	0.87
MI vs. Super-diagnostic classes	91.07%	90.2%	0.762

TABLE 11. Five Fold Cross-Validations.

Data	Alg.	Fold 0	1	2	3	4	Avg
ECG HAR	1	97.26	98.43	98.43	98.03	98.82	98.19
	2	94.90	90.59	91.37	93.70	93.31	92.77
PTB	1	99.00	99.56	99.28	99.07	99.31	99.24
	2	96.39	97.11	97.49	96.25	96.49	96.75
PTB XL	1	76.22	75.42	75.38	74.00	76.49	75.50

D. STATISTICAL ANALYSIS

To assess the statistical validity of the results, we computed a five-fold cross-validation for both algorithms and all data sets, see Table 11. For PTB-XL data set the k-fold validation was done only for super-diagnostic classes.

As one can see, the results are consistent with the findings in sub section VI-A, the difference in the average results from the reduced training in case of five fold cross validation.

However, to assess the statistical meaningfulness, a more sophisticated approach is required. In a seminal paper [60], Dietterich analyzed five approximate statistical tests for determining whether one learning algorithm outperforms another on a particular learning task. It includes the well-known McNemar's test for a single pass validation and also proposes a new 5×2cv test designed for algorithms where at least 10 validations can be carried out. In the paper, Dietterich shows that the null-hypothesis of the two algorithms to compare having the same performance, the off-diagonal elements of the confusion matrix should be the same, which can be checked statistically for significance using a χ^2 test or a test for t statistics.

Although Dietterich's paper has been very well received and is cited many thousand times, in the practice of machine learning–contrary to other disciplines like e.g., the medical sciences–, statistical analysis of significance is still not common and is therefore usually not included in publications, unfortunately. This limits not only the interpretation of published results but also limits the ability to rigorously compare benchmarks. For instance, the tests proposed in [60] assume the availability of the data set backing the confusion matrix. In particular, to apply any of the tests in order to compare two algorithms A_1 and A_2, one must determine the incorrectly classified samples from A_1 and check whether these are correctly classified by algorithm A_2 and vice versa in order to determine the statistical parameters needed for the test. Even if the data are publicly available, information on which sub samples are incorrectly classified is usually *not* available from the publication. In our case, these data are clearly not available for any of the bench marking publications. Therefore, we could only compare our algorithms 1 and 2 against each other, but *not* against any of the bench marked algorithms.

The implementation of the statistical analysis and the results are discussed in the next sub section.

1) McNemar's TEST

The McNemar's test is a standard paired test used in the medical field for the verification of usability of the new drugs etc. However, it is not very commonly applied in the field of deep learning for model comparison. Because we used biomedical data in this study, McNemar's test was used to verify the statistical significance of the results obtained using the proposed algorithms. To apply McNemar's test, our data

TABLE 12. McNemar's Contingency Tables.

n_{00}	n_{10}
n_{01}	n_{11}

(a) Layout

6	31
1	217

(b) ECG Data Set

1	11
4	1440

(c) PTB Data Set

set was partitioned into training and testing sets, called R and T, respectively. After training both models A_1 (Algorithm 1) and A_2 (Algorithm 2) on t, the classifiers were tested on each instance of R and eventually the following statistics are collected:

- n_{00}: Number of classes misclassified by both classifiers A_1 and A_2.
- n_{01}: Number of data instances misclassified by A_1 but not A_2.
- n_{10}: Number of data instances misclassified by A_2 but not A_1.
- n_{11}: Number of data instances misclassified by neither A_1 nor A_2.

Additionally, $n_{00}+n_{01}+n_{10}+n_{11} = n$, where n is the total data instances in the test set T. The contingency table layout for McNemar's test is presented in Table 12a.

The McNemar's test was performed for both PTB and ECG HAR data set. The null hypothesis (H_0) is that both algorithms have the same error rate, i.e., $n_{01} = n_{10}$. The confidence interval for all tests was 95%. The statistics obtained for the ECG HAR data set are presented in Table 12b.

For an alpha value of 0.05, the p-value is calculated to be numerically 0.000, which implies that our test is significant enough to reject the H_0 and we conclude that both models have different proportions of errors and are significantly different in this data set. The same test was repeated for the PTB data set and the obtained statistics are listed in Table 12c.

The p-value obtained for this test was 0.118 which is greater than 0.05 hence, there was no significant evidence to reject H_0.

2) THE 5 × 2 CV t TEST

The 5 × 2 cv t test is introduced in Dietterich [60] and recommended therein: "For algorithms that can be executed ten times, the 5 × 2 cv test is recommended as it is slightly more powerful and because it directly measures the variance due to the choice of training set". For this test, two-fold cross validation was performed for five repetitions. During every repetition, the data set was randomly partitioned into two equal-sized sets S_1 and S_2. Both algorithms were trained on each set and tested on the other set. This results in four error estimates: $P_{A_1}^{(1,2)}$ and $P_{A_2}^{(1,2)}$ with A_1 or A_2, resp. trained on S_1 and tested on S_2 and $P_{A_1}^{(2,1)}$ and $P_{A_2}^{(2,1)}$ with A_1 or A_2, resp. trained on S_2 and tested on S_1. Estimated differences are obtained by subtracting the corresponding error estimates $P^{(1,2)} = P_{A_1}^{(1,2)} - P_{A_2}^{(1,2)}$ and $P^{(2,1)} = P_{A_1}^{(2,1)} - P_{A_2}^{(2,1)}$. From these differences, the estimated variance σ^2 is calculated as $\sigma^2 = (P^{(1,2)} - \bar{P})^2 + (P^{(2,1)} - \bar{P})^2$, where $\bar{P} = (P^{(1,2)} + P^{(2,1)})/2$. Let σ_i^2 be the variance calculated from the i-th replication. Then the 5 × 2cv $\bar{t}$ statistic is calculated as follows:

$$\bar{t} = \frac{P_1^{(1,2)}}{\sqrt{\frac{1}{5}\sum_{i=1}^{5}\sigma_i^2}}$$

Under the H_0, $\bar{t}$ has approximately a t distribution with five degrees of freedom. The calculated t statistic for PTB data set and ECG HAR data set is 3.002 and 3.286 respectively. Detailed tables for the five repetitions are presented in Table 15 and Table 16. As both would have a corresponding p value of 0.030 and 0.0218 respectively, it clearly shows that both models are significantly different from each other with different error estimates.

3) INTERPRETATION

Taking the results from both the McNemar and the more powerful 5 × 2 cv t test we can conclude that our algorithms differ significantly and that the obtained Key performance Indicators (KPIs) are statistically meaningful for the standard confidence interval of 95%.

E. SUMMARY

As seen in Table 13, our model leads to almost all of the evaluation criteria for the classification of PTB data set.

VII. DISCUSSION

As mentioned earlier, state-of-the-art accuracies were achieved using the CNN-LSTM model for three data sets and the attention model for PTB data set. In similar previous studies, mostly one set of experiments is performed with a single database to prove the usability of the models. However, we worked on three data sets separately. The first data set was used to classify human activities including falls. The second one consists of extracted heartbeats for the classification of MI vs normal heartbeats. The third data set consists of a 12-lead ECG data set for multiple cardiovascular conditions. The success of our proposed algorithms on all three data sets generalizes their usefulness for ECG classifications over multiple tasks.

Hybrid models help to combine the features of the base models. This is often more powerful than very deep models with hundreds of layers because deeper models tend to over-fit for medium-sized data sets. An LSTM model keeps track of the past trends in the time series and can also help in the prediction of the next time stamps. In our study, the results of the CNN-LSTM model have shown to be always better than both of the models implemented individually. This was verified for the HAR data set by [61] and we compare the results from Table 13 for PTB data set where multiple variations of CNN and LSTMs have been applied separately in the previous works. The performance of the model on the HAR data set is observed to increase up to a certain level with the increase in a) the number of filters in the conv1d

TABLE 13. Our Result Compared with other similar Studies in Literature which used PTB Database (Built upon [49]).

Work	Accuracy(%)	Sensitivity(%)	Specificity(%)	Precision(%)
Acharya et al. [52]	93.5	93.7	-	92.8
Safdarian et al. [62]	94.7	-	-	-
Kojuri et al. [63]	95.6	93.3	-	97.9
Sun et al. [64]	-	92.6	-	82.4
Liu et al. [65]	94.4	-	-	-
Sharma et al. [66]	96	93	-	99
Kachuee et al. [18]	95.9	95.1	-	95.2
Remya et al. [67]	93.61	93.22	94.28	-
Reasat et al. [68]	84.54	85.33	84.09	-
Zewdie et al. [69]	98.3	98.7	96.4	-
Feng et al. [59]	95.4	98.2	86.5	-
Strodthoff et al.	-	93.3	89.7	-
Huang et al.	96.96	99.89	92.51	95.35
Liu et al. [58]	98.59	99.53	94.50	-
Gupta et al. [49]	99.43	99.40	99.45	99.46
Ours (CNN-LSTM)	**99.93**	**99.81**	99.73	**99.91**
Ours (Attention)	**99.73**	99.72	**99.73**	**99.91**

TABLE 14. Overview of the Experiments with Different Data Sets and the Acquired Performances.

Classification task	Data set	Achieved Accuracy	Algorithm	State-of-the-art accuracy
Fall detection	ECG HAR	99.21%	Attention	98.44% [7]
Fall detection	ECG HAR	99.21%	CNN-LSTM	98.44% [7]
MI detection	PTB	99.73%	Attention	99.44% [49]
MI detection	PTB	99.93%	CNN-LSTM	99.44% [49]
SD Class	PTB XL	75.70%	CNN-LSTM	-
MI detection	PTB XL	91.07%	CNN-LSTM	-

TABLE 15. 5 × 2 cv Test Contingency Table for PTB Data Set.

	Rep 1	Rep 2	Rep 3	Rep 4	Rep 5
Model A	$P_A^{(1)} = 0.9915$ $P_A^{(2)} = 0.9922$	$P_A^{(1)} = 0.9923$ $P_A^{(2)} = 0.9904$	$P_A^{(1)} = 0.9934$ $P_A^{(2)} = 0.9889$	$P_A^{(1)} = 0.9934$ $P_A^{(2)} = 0.9927$	$P_A^{(1)} = 0.9929$ $P_A^{(2)} = 0.9894$
Model B	$P_B^{(1)} = 0.9786$ $P_B^{(2)} = 0.9839$	$P_B^{(1)} = 0.9839$ $P_B^{(2)} = 0.9733$	$P_B^{(1)} = 0.9778$ $P_B^{(2)} = 0.9812$	$P_B^{(1)} = 0.9799$ $P_B^{(2)} = 0.9805$	$P_B^{(1)} = 0.9759$ $P_B^{(2)} = 0.9789$

TABLE 16. 5 × 2 cv Test Contingency Table for ECG Data Set.

	Rep 1	Rep 2	Rep 3	Rep 4	Rep 5
Model A	$P_A^{(1)} = 0.96860$ $P_A^{(2)} = 0.96698$	$P_A^{(1)} = 0.95918$ $P_A^{(2)} = 0.977987$	$P_A^{(1)} = 0.97327$ $P_A^{(2)} = 0.96232$	$P_A^{(1)} = 0.9545$ $P_A^{(2)} = 0.9733$	$P_A^{(1)} = 0.96232$ $P_A^{(2)} = 0.9733$
Model B	$P_B^{(1)} = 0.86656$ $P_B^{(2)} = 0.8522$	$P_B^{(1)} = 0.8477$ $P_B^{(2)} = 0.8805$	$P_B^{(1)} = 0.8349$ $P_B^{(2)} = 0.85714$	$P_B^{(1)} = 0.85714$ $P_B^{(2)} = 0.88522$	$P_B^{(1)} = 0.85714$ $P_B^{(2)} = 0.8349$

layer for CNN-LSTM and b) the number of dimensions in the dimensional embedding with the attention model. Since the data set is not very large, a final conclusion cannot be drawn at this stage but it merits further investigation. The attention algorithm clearly has a computational advantage over the CNN-LSTM algorithm as seen in Table 8. It takes less time to converge and even has fewer parameters to train than the CNN-LSTM algorithm. Our study had the following advantages:

- A hybrid CNN-LSTM model and attention with a discrete wavelet transformation as an embedding are proposed.
- No or very little manual feature extraction is required for training the model.
- Three publicly available data sets were used separately for the training using the proposed models.
- State-of-the-art accuracy of 99.86% and 99.44% is achieved for the PTB data set and ECG for HAR

classification respectively without any feature extraction or pre-processing.
- Multiple standard statistical analysis techniques were applied to the acquired results to statistically support our algorithms.

Hence, we addressed our research question and achieved results equivalent to many recent studies without any pre-processing or feature extraction. We have also shown to train the models in an efficient manner computationally.

As part of the future work, the authors would like to explore the difference between the two algorithms using explainable AI. Looking deeper into the gradients for each layer would shed light into the learning process.

VIII. CONCLUSION

The models proposed and explained in this paper aim to better classify ECG time series for different conditions using minimum pre-processing steps. Publicly available data sets have made it possible to verify the robustness and usefulness of the proposed models by achieving state-of-the-art accuracy using multiple data sets. This would eventually help medical practitioners to identify multiple heart conditions automatically with minimum feature extraction. Specifically for the MI classification, because the results are close to 100%, the model is ready to be deployed for medical evaluation.

APPENDIX

DATA AND CODE AVAILABILITY

The data sets used are publicly available and present in the corresponding repositories:
- ECG HAR data set: [70]
- PTB DB data set: [48]
- PTB XL: [6]

The code is available on GitHub at https://github.com/buttfatimasajid/Towards-Automated-Feature-Extraction-For-Deep-Learning-Classification-of-Electrocardiogram-Signals.

ACKNOWLEDGMENT

The authors thank Kylie Pusch who contributed some ideas in their bachelor thesis [61].

REFERENCES

[1] *Centers for Disease Control and & Prevention*. Underlying Cause of Death, 1999–2020 Request. Accessed: Mar. 5, 2022. [Online]. Available: https://wonder.cdc.gov/ucd-icd10.html
[2] M. Tan and R. Kenny, "Cardiovascular assessment of falls in older people," *Clin. Intervent. Aging*, vol. 1, no. 1, pp. 57–66, 2006.
[3] S. Karpagachelvi, M. Arthanari, and M. Sivakumar, "ECG feature extraction techniques—A survey approach," 2010, *arXiv:1005.0957*.
[4] S. H. Jambukia, V. K. Dabhi, and H. B. Prajapati, "Classification of ECG signals using machine learning techniques: A survey," in *Proc. Int. Conf. Adv. Comput. Eng. Appl.*, Mar. 2015, pp. 714–721.
[5] M. Ingale, R. Cordeiro, S. Thentu, Y. Park, and N. Karimian, "ECG biometric authentication: A comparative analysis," *IEEE Access*, vol. 8, pp. 117853–117866, 2020.
[6] P. Wagner, N. Strodthoff, R.-D. Bousseljot, D. Kreiseler, F. I. Lunze, W. Samek, and T. Schaeffter, "PTB-XL, a large publicly available electrocardiography dataset," *Sci. Data*, vol. 7, no. 1, pp. 1–15, Dec. 2020, doi: 10.1038/s41597-020-0495-6.
[7] F. S. Butt, L. La Blunda, M. F. Wagner, J. Schäfer, I. Medina-Bulo, and D. Gómez-Ullate, "Fall detection from electrocardiogram (ECG) signals and classification by deep transfer learning," *Information*, vol. 12, no. 2, p. 63, Feb. 2021, doi: 10.3390/info12020063.
[8] H. Khorrami and M. Moavenian, "A comparative study of DWT, CWT and DCT transformations in ECG arrhythmias classification," *Exp. Syst. Appl.*, vol. 37, no. 8, pp. 5751–5757, Aug. 2010.
[9] L. Sathyapriya, L. Murali, and T. Manigandan, "Analysis and detection R-Peak detection using modified Pan-Tompkins algorithm," in *Proc. IEEE Int. Conf. Adv. Commun., Control Comput. Technol.*, May 2014, pp. 483–487.
[10] A. Dallali, A. M. Kachouri, and A Samet, "Classification of cardiac arrhythmia using WT, HRV, and fuzzy C-means clustering," *Signal Process., Int. J.*, vol. 5, no. 3, pp. 101–108, 2011.
[11] Z. Ebrahimi, M. Loni, M. Daneshtalab, and A. Gharehbaghi, "A review on deep learning methods for ECG arrhythmia classification," *Exp. Syst. Appl., X*, vol. 7, Sep. 2020, Art. no. 100033. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2590188520300123
[12] R. Socher, B. Huval, B. Bath, C. A. Manning, and A. Ng, "Convolutional-recursive deep learning for 3D object classification," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1–9. [Online]. Available: https://proceedings.neurips.cc/paper/2012/file/3eae62bba9ddf64f69d49dc48e2dd214-Paper.pdf
[13] Y. Zheng, Q. Liu, E. Chen, Y. Ge, and J. L. Zhao, *Time Series Classification Using Multi-Channels Deep Convolutional Neural Networks* (Lecture Notes in Computer Science), vol. 8485, F. Li, G. Li, S. Hwang, B. Yao, and Z. Zhang, Eds. Cham, Switzerland: Springer, doi: 10.1007/978-3-319-08010-9_33.
[14] K. Xia, J. Huang, and H. Wang, "LSTM-CNN architecture for human activity recognition," *IEEE Access*, vol. 8, pp. 56855–56866, 2020, doi: 10.1109/ACCESS.2020.2982225.
[15] F. Ordóñez and D. Roggen, "Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, p. 115, Jan. 2016, doi: 10.3390/s16010115.
[16] Ö. Yildirim, P. Plawiak, R.-S. Tan, and U. R. Acharya, "Arrhythmia detection using deep convolutional neural network with long duration ECG signals," *Comput. Biol. Med.*, vol. 102, pp. 411–420, Nov. 2018, doi: 10.1016/j.compbiomed.2018.09.009.
[17] Y. H. Awni, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng, "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nature Med.*, vol. 25, pp. 65–69, Jan. 2019.
[18] M. Kachuee, S. Fazeli, and M. Sarrafzadeh, "ECG heartbeat classification: A deep transferable representation," in *Proc. IEEE Int. Conf. Healthcare Informat. (ICHI)*, Jun. 2018, pp. 443–444.
[19] N. Strodthoff and C. Strodthoff, "Detecting and interpreting myocardial infarction using fully convolutional neural networks," *Physiolog. Meas.*, vol. 40, no. 1, Jan. 2019, Art. no. 015001, doi: 10.1088/1361-6579/aaf34d.
[20] J. Wang, X. Qiao, C. Liu, X. Wang, Y. Liu, L. Yao, and H. Zhang, "Automated ECG classification using a non-local convolutional block attention module," *Comput. Methods Programs Biomed.*, vol. 203, May 2021, Art. no. 106006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016926072100081X
[21] T. Wang, C. Lu, Y. Sun, M. Yang, C. Liu, and C. Ou, "Automatic ECG classification using continuous wavelet transform and convolutional neural network," *Entropy*, vol. 23, no. 1, p. 119, Jan. 2021. [Online]. Available: https://www.mdpi.com/1099-4300/23/1/119
[22] S. Saadatnejad, M. Oveisi, and M. Hashemi, "LSTM-based ECG classification for continuous monitoring on personal wearable devices," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 2, pp. 515–523, Feb. 2020, doi: 10.1109/JBHI.2019.2911367.
[23] Z. Gao, X. Wang, S. Sun, D. Wu, J. Bai, Y. Yin, X. Liu, H. Zhang, and V. H. C. de Albuquerque, "Learning physical properties in complex visual scenes: An intelligent machine for perceiving blood flow dynamics from static CT angiography imaging," *Neural Netw.*, vol. 123, pp. 82–93, Mar. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0893608019303764
[24] S. M. Mathews, C. Kambhamettu, and K. E. Barner, "A novel application of deep learning for single-lead ECG classification," *Comput. Biol. Med.*, vol. 99, pp. 53–62, Aug. 2018, doi: 10.1016/j.compbiomed.2018.05.013.
[25] G. Litjens, F. Ciompi, J. M. Wolterink, B. D. de Vos, T. Leiner, J. Teuwen, and I. Isgum, "State-of-the-art deep learning in cardiovascular image analysis," *JACC: Cardiovascular Imag.*, vol. 12, no. 8, pp. 1549–1565, Aug. 2019, doi: 10.1016/j.jcmg.2019.06.009.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. I. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[27] H. Ramsauer, B. Schäfl, J. Lehner, P. Seidl, M. Widrich, T. Adler, L. Gruber, M. Holzleitner, M. Pavlovic, G. Sandve, V. Greiff, D. Kreil, M. Kopp, G. Klambauer, and J. S. Brandstetter, "Hopfield networks is all you need," 2020, *arXiv:2008.02217*.

[28] G. Lai, W.-C. Chang, Y. Yang, and H. Liu, "Modeling long- and short-term temporal patterns with deep neural networks," 2017, *arXiv:1703.07015*.

[29] S.-Y. Shih, F.-K. Sun, and H.-Y. Lee, "Temporal pattern attention for multivariate time series forecasting," *Mach. Learn.*, vol. 108, nos. 8–9, pp. 1421–1441, Sep. 2019. [Online]. Available: http://dblp.uni-trier.de/db/journals/ml/ml108.html#ShihSL19

[30] H. Song, D. Rajan, J. A. Thiagarajan, and A. Spanias, "Attend and diagnose: Clinical time series analysis using attention models," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 4091–4098.

[31] G. Yan, S. Liang, Y. Zhang, and F. Liu, "Fusing transformer model with temporal features for ECG heartbeat classification," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Nov. 2019, pp. 898–905.

[32] S. Wang, B. Z. Li, M. Khabsa, H. Fang, and H. Ma, "Linformer: Self-attention with linear complexity," 2020, *arXiv:2006.04768*.

[33] M. Rabe and C. Staats, "Self-attention does not need $O(n^2)$ memory," 2021, *arXiv:2112.05682*.

[34] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," *Artif. Intell. Rev.*, vol. 53, no. 8, pp. 5455–5516, Dec. 2020, doi: 10.1007/s10462-020-09825-6.

[35] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in *Proc. Int. Conf. Eng. Technol. (ICET)*, Aug. 2017, pp. 1–6.

[36] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[37] S. M. Mathews, "Dictionary and deep learning algorithms with applications to remote health monitoring systems," Ph.D. thesis, Dept. Elect. Comput. Eng., Univ. Delaware, Newark, DE, USA, 2017. [Online]. Available: http://udspace.udel.edu/handle/19716/21241

[38] I. Daubechies, *Ten Lectures on Wavelets* (Society for Industrial). Philadelphia, PA, USA: SIAM, Aug. 1992.

[39] S. Mallat and S. Zhong, "Characterization of signals from multiscale edges," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 7, pp. 710–732, Jul. 1992.

[40] Ø. Ryan and L. Algebra, *Signal Processing, and Wavelets—A Unified Approach: Python Version*. Cham, Switzerland: Springer, 2019.

[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[42] V. Strassen, "Gaussian elimination is not optimal," *Numer. Math.*, vol. 13, no. 4, pp. 354–356, 1969, doi: 10.1007/BF02165411.

[43] J. Alman and V. V. Williams, "A refined laser method and faster matrix multiplication," 2020, *arXiv:2010.05846*.

[44] L. L. Blunda, L. Gutierrez-Madronal, M. F. Wagner, and I. Medina-Bulo, "A wearable fall detection system based on body area networks," *IEEE Access*, vol. 8, pp. 193060–193074, 2020.

[45] R. Kher, "Wearable ambulatory electrocardiogram (ECG) and EEG dataset," *IEEE Dataport*, 2020, doi: 10.21227/ysnc-gc65.

[46] Y.-G. Kim, D. Shin, M. Y. Park, S. Lee, M. S. Jeon, D. Yoon, and R. W. Park, "ECG-ViEW II, a freely accessible electrocardiogram database," *PLoS ONE*, vol. 12, no. 4, Apr. 2017, Art. no. e0176222. [Online]. Available: https://europepmc.org/articles/PMC5402933

[47] Z. Cui, W. Chen, and Y. Chen, "Multi-scale convolutional neural networks for time series classification," 2016, *arXiv:1603.06995*.

[48] R. Bousseljot, D. Kreiseler, and A. Schnabel, "Nutzung der EKG-Signaldatenbank CARDIODAT der PTB uber das internet," *Biomedizinische Technik, Band*, vol. 40. Berlin, Germany: German Heart Institute, Jan. 1995, doi: 10.13026/C28C71.

[49] A. Gupta, E. Huerta, Z. I. Zhao, and I. Moussa, "Deep learning for cardiologist-level myocardial infarction detection in electrocardiograms," in *Proc. 8th Eur. Med. Biol. Eng. Conf.*, 2021, pp. 341–355.

[50] N. Strodthoff, P. Wagner, T. Schaeffter, and W. Samek, "Deep learning for ECG analysis: Benchmarks and insights from PTB-XL," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 5, pp. 1519–1528, May 2021.

[51] F. Yang, G. Wang, C. Luo, and Z. Ding, "Improving automatic detection of ECG abnormality with less manual annotations using Siamese network," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Nov. 2021, pp. 1120–1123.

[52] U. R. Acharya, H. Fujita, S. L. Oh, Y. Hagiwara, J. H. Tan, and M. Adam, "Application of deep convolutional neural network for automated detection of myocardial infarction using ECG signals," *Inf. Sci.*, vols. 415–416, pp. 190–198, Nov. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025517308009

[53] H. M. Rai and K. Chatterjee, "Hybrid CNN-LSTM deep learning model and ensemble technique for automatic detection of myocardial infarction using big ECG data," *Int. J. Speech Technol.*, vol. 52, no. 5, pp. 5366–5384, Mar. 2022, doi: 10.1007/s10489-021-02696-6.

[54] E. Ramaraj, "A novel deep learning based gated recurrent unit with extreme learning machine for electrocardiogram (ECG) signal recognition," *Biomed. Signal Process. Control*, vol. 68, Jul. 2021, Art. no. 102779. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809421003761

[55] H. Martin, U. Morar, W. Izquierdo, M. Cabrerizo, A. Cabrera, and M. Adjouadi, "Real-time frequency-independent single-lead and single-beat myocardial infarction detection," *Artif. Intell. Med.*, vol. 121, Nov. 2021, Art. no. 102179. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S093336572100172X

[56] S. Smigiel, K. Palczynski, and D. Ledzinski, "Deep learning techniques in the classification of ECG signals using R-Peak detection based on the PTB-XL dataset," *Sensors*, vol. 21, no. 24, p. 8174, Dec. 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/24/8174

[57] J. Schäfer, "Human activity recognition with CSI data—Attention is all you need," Fac. 2 Comput. Sci. Eng., Frankfurt Univ. Appl. Sci., Frankfurt, Germany, Tech. Rep., 2022.

[58] N. Liu, L. Wang, Q. Chang, Y. Xing, and X. Zhou, "A simple and effective method for detecting myocardial infarction based on deep convolutional neural network," *J. Med. Imag. Health Informat.*, vol. 8, no. 7, pp. 1508–1512, Sep. 2018.

[59] K. Feng, X. Pi, H. Liu, and K. Sun, "Myocardial infarction classification based on convolutional neural network and recurrent neural network," *Appl. Sci.*, vol. 9, no. 9, p. 1879, May 2019.

[60] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Comput.*, vol. 10, no. 7, pp. 1895–1923, Oct. 1998, doi: 10.1162/089976698300017197.

[61] K. Pusch, "ECG classification using different machine learning models for human activity recognition," Bachelor thesis, Fac. 2 Comput. Sci. Eng., Frankfurt Univ. Appl. Sci., Frankfurt, Germany, 2021.

[62] N. Safdarian, N. J. Dabanloo, and G. Attarodi, "A new pattern recognition method for detection and localization of myocardial infarction using T-wave integral and total integral as extracted features from one cycle of ECG signal," *J. Biomed. Sci. Eng.*, vol. 7, no. 10, pp. 818–824, 2014, doi: 10.4236/jbise.2014.710081.

[63] J. Kojuri, R. Boostani, P. Dehghani, F. Nowroozipour, and N. Saki, "Prediction of acute myocardial infarction with artificial neural networks in patients with nondiagnostic electrocardiogram," *J. Cardiovascular Disease Res.*, vol. 6, no. 2, pp. 51–59, May 2015.

[64] L. Sun, Y. Lu, K. Yang, and S. Li, "ECG analysis using multiple instance learning for myocardial infarction detection," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 12, pp. 3348–3356, Dec. 2012.

[65] B. Liu, J. Liu, G. Wang, K. Huang, F. Li, Y. Zheng, Y. Luo, and F. Zhou, "A novel electrocardiogram parameterization algorithm and its application in myocardial infarction detection," *Comput. Biol. Med.*, vol. 61, pp. 178–184, Jun. 2015.

[66] L. Sharma and R. Sunkaria, "Inferior myocardial infarction detection using stationary wavelet transform and machine learning approach," *Signal, Image Video Process.*, vol. 12, no. 2, pp. 199–206, 2018.

[67] R. S. Remya, K. P. Indiradevi, and K. K. A. Babu, "Classification of myocardial infarction using multi resolution wavelet analysis of ECG," *Proc. Technol.*, vol. 24, pp. 949–956, Jan. 2016.

[68] R. Tahsin and C. Shahnaz, "Detection of inferior myocardial infarction using shallow convolutional neural networks," in *Proc. 7th IEEE Region Humanitarian Technol. Conf.*, Dec. 2017, pp. 718–721.

[69] G. Zewdie and M. Xiong, "Fully automated myocardial infarction classification using ordinary differential equations," 2014, *arXiv:1410.6984*.

[70] F. Butt, L. La Blunda, M. Wagner, J. Schäfer, and I. D. Medina-Bulo, "ECG data for deep transfer learning," *IEEE Dataport*, 2020, doi: 10.3390/info12020063.

FATIMA SAJID BUTT received the bachelor's degree in information technology from the University of the Punjab, Lahore, Pakistan, in 2010, and the master's degree in high integrity systems (HIS) from the Frankfurt University of Applied Sciences, Frankfurt, Germany, in 2019. She is currently pursuing the doctoral degree in engineering informatics with the University of Cádiz, Spain. She is a Research Assistant with the Frankfurt University of Applied Sciences. She is a part of the Research Group Industrial Data Sciences (INDAS) along with Jörg Schäfer, Matthias Wagner and Dirk Stegelmeyer, Frankfurt University of Applied Sciences. Her research interests include time series analysis for classification and application of machine learning algorithms for industrial problems, such as predictive maintenance.

MATTHIAS F. WAGNER (Member, IEEE) received the Diploma and Dr.rer.nat. degrees in physics from the Johannes Gutenberg-Universität Mainz, Germany. He was the Head of Measuring Technology Software Development at Hottinger Baldwin Messtechnik (HBM), Darmstadt, Germany, from 1990 to 2002. In 2002, he was appointed as a Professor of computer science with the Frankfurt University of Applied Sciences, Frankfurt am Main, Germany. Since 2005, he has been the Program Director of the international M.Sc. program ''High Integrity Systems.'' From 2017 to March 2020, he served as the Vice-Dean for Research and International Relations of the FB2, Department of Computer Science and Engineering. Since 2010, he has been the Head of the Research Group Wireless Sensor Networks and Internet of Things (WSN and IoT). He was supported by research stays at the UCASE Software Engineering Research Group, Universidad de Cádiz, Spain, and the Dipartimento di Fisica e Astronomia, Università degli Studi di Firenze, Italy. His research interests include safety critical computer systems, smart sensors and actuator networks, software and systems engineering and computational science.

JÖRG SCHÄFER (Member, IEEE) received the Ph.D. degree in mathematical physics from Bochum University, Germany, in 1992. After spending more than ten years working as the Principal Architect in IT consulting for large international companies he was appointed as a Full Professor of computer science with the Frankfurt University of Applied Sciences, in 2009. Since 2012, he has been the Chair of the computer science B.Sc. program. His main research interests include theoretical understanding of deep learning architectures and applying machine learning and probabilistic models in ubiquitous computing applications. He runs the Research Group on Human Activity Recognition (HAR) and Channel State Information (CSI) and jointly with Matthias Wagner and Dirk Stegelmeyer the Research Group Industrial Data Science (INDAS). He is collaborating with UCASE Software Engineering Research Group of the Universidad de Cádiz, Spain.

DAVID GOMEZ ULLATE received the Ph.D. degree in physics from Universidad Complutense de Madrid, in 2001. He is a Distinguished Researcher with University of Cádiz, where he founded and serves as the Director of UCA Datalab. He is a Professor of applied mathematics on leave from the Complutense University of Madrid, an Adjunct Professor with IE Business School, and a Visiting Professor of mathematics and data science with the University of Loughborough, U.K. His major contribution is the theory of exceptional orthogonal polynomials, which has earned him international recognition expressed as plenary talks in the main conferences of the field, and invited seminars at Cambridge, Harvard, Edinburgh, Stockholm, Copenhagen, and Rome, among others. For the past ten years he has specialized in knowledge transfer of mathematics to industry. His research interests include mathematical physics, approximation theory, applied machine learning, and data science. He is currently the President of the Knowledge Transfer Commission of the Royal Spanish Mathematical Society, and a member of its Governing Board. In 2016, he received the Leonardo Scholarship from the BBVA Foundation for a project on credit card fraud detection. He has coordinated eight knowledge transfer contracts with industry in the financial, fisheries, biomedical, legaltech and aeronautical sectors. He is a Co-Founder and a Scientific Advisor of the technological startup Komorebi AI.