Classification of 27 heart abnormalities using 12-lead ECG signals with combined deep learning techniques

Atiaf A. Rawi[1], Murtada Khalafallah Elbashir[2], Awadallah M. Ahmed[1]
[1]Department of Computer Sciences, Faculty of Mathematical and Computer Sciences, Gezira University, Wad Madani, Sudan
[2]Department of Information Systems, College of Computer and Information Sciences, Jouf University, Sakaka, Saudi Arabia

Article Info	ABSTRACT
	An electrocardiogram (ECG) machine with a standard 12-lead configuration is the primary clinical technique for diagnosing abnormalities in heart function. Automated 12-lead ECG machines have the capacity to screen the general population and provide second opinions for physicians. However, expertise and time are required for manual ECG interpretation. Therefore, computer-aided diagnoses are of interest to the medical community. Hence, this study aims to build a deep learning (DL) model with an end-to-end structure that can categorize 12-lead ECG results into 27 different disorders. We use multivariate time-series data to construct a novel end-to-end DL model (based on combined convolutional neural networks (CNNs), long short-term memory, gated recurrent units, and a deep residual network structure) for feature representations and determining spatial relations among deep features. In addition, a dataset of 43,101 classified standard ECG recordings was collected from six different sources to guarantee the model's ability to generalize and alleviate data divergence. As a result, the residual network-based model obtained promising outcomes and an accuracy of 0.97. According to the experimental data, it outperforms other methods. *This is an open access article under the [CC BY-SA](#) license.*
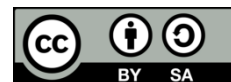
Corresponding Author:

Atiaf A. Rawi
Department of Computer Sciences, Faculty of Mathematical and Computer Sciences, Gezira University
Wad Madani, Sudan
Email: atiaf.ayal88@gmail.com

1. INTRODUCTION

According to the World Health Organization (WHO), worldwide, cardiovascular diseases (CVDs) are the leading cause of death, killing 18 million people annually. CVDs include coronary heart disease, cerebrovascular disease, rheumatic heart disease, and various heart and blood vessel issues. Heart attacks and strokes account for more than four in every five CVD deaths, with one-third occurring before age 70. However, treatment costs and sudden cardiac deaths can be reduced significantly with the help of accurate and early diagnoses [1]. Electrocardiogram (ECG) machines are widely used for CVD diagnosis due to their inexpensiveness, high accuracy, and non-invasive nature. They use 12 electrocardiograph leads to record the heart's electrical activity.

The resulting sequence of electrical signals is recorded from different places on the human body [2]. However, skilled doctors are required to investigate and identify abnormal inter- and intra-beat patterns picked up by an ECG. Moreover, this process is time-consuming and vulnerable to inter-observer variability [3], making an automated ECG signal classification system essential, particularly in non-cardiology departments and pre-hospital care settings, where an expert may not always be accessible to interpret ECG signals [4]. Many of the earlier methods for automated ECG signal analysis were based on signal transformation (such as fourier and wavelet), time-frequency, and frequency domain features [5]-[7].

However, in addition to their complexity, they were unable to capture complex features in ECG data. Recently, artificial intelligence (AI) and deep learning (DL) algorithms have been developed to process large-scale raw data, avoiding hand-crafted feature extraction methods [8]. Convolutional neural networks (CNNs) have achieved notable success in many fields, such as natural language processing [9] and computer vision [10]. These successes motivated researchers, as in [11], to propose a multi-layer 1-D recurrent neural network (RNN) trained on ECG data from a single lead. A CNN-based method has been suggested to increase classification accuracy [12]. In addition, two deep network models using short single-lead signals have been proposed for classifying pulse-generating and pulse-less rhythmic activities [13]. Another single-lead-based method has been presented that features an ensemble DL model for automating ECG signal classification [14]. Indexing and abstracting services depend on the accuracy of the title, extracting from it keywords useful in cross-referencing and computer searching. An improperly titled paper may never reach the audience for which it was intended, so be specific.

In this method, ten classifiers are fused, and it produced better results than the single, profound classifiers. A multi-stage learning model introduced features such as the frequency and rhythm of beats [15]. Inspired by the performances achieved by previous neural network-based methods, which proved their ability to capture nonlinearity and complex features, our proposed method is also based on a neural network approach. In most past approaches, two to nine heart abnormalities have been classified. Furthermore, most of the currently available methods handle ECG data from a single lead, even though 12-lead data are more widely used in real-life diagnostic settings. Additionally, most of these works treated diagnosis as a multi-class classification problem, while multiple abnormalities often appear in the same ECG record. Some proposed methods for analyzing 12-lead ECG data efficiently for 27 heart abnormalities treat this problem as a multi-label classification problem, allowing them to consider the presence of more than one abnormality simultaneously. The main contributions of this work can be summarized as follows: i) it presents an end-to-end model for classifying 27 heart abnormalities using 12-lead ECG signals, ii) we combined techniques to enhance feature extraction and classification accuracy, iii) extensive experiments were conducted on combined datasets from six different sources to ensure that the model was generalizable, iv) the performance comparison proved that the suggested method outperformed the state-of-the-art methods without the need for pre-processing or manual feature engineering, and v) the suggested method treats the classification problem as a multi-label classification to handle multiple abnormalities in the same ECG record.

Data to classify heart abnormalities. Chen et al. [16] used a ResNet [17] structure with 1-D convolutional layers for feature extraction; their network outputs a 1×512 vector for each lead. By concatenating the 12 resulting vectors, a matrix of size 12×512 is obtained, which is then fed into a long-short term memory (LSTM) layer and a fully connected layer for final classification. This method classified seven heart abnormalities and trained on 7,704 samples. Liu et al. [18] used a biorthogonal wavelet transformation to denoise ECG signals. Then the E-ResNet [19] model was used as a baseline model. Furthermore, Pan and Tompkin's algorithm [20] was applied to detect the R-peaks on lead Baloglu et al. [21] proposed a deep CNN model with ten layers to classify 11 classes of 651 samples each. The suggested model was trained for each lead signal individually. Fayyazifar et al. [22] suggested a model of 49 1-D CNN layers, one LSTM layer, and 16 skip connections to classify 27 ECG signal types. Leur et al. [8] utilized the 10-second raw data of 8-lead signals (I, II, and V1–V6), sampled at 500 Hz, as the input for a deep neural network with a structure similar to that of the Inception ResNet network by combining convolutional layers with skip connections in parallel.

Gliner et al. [23] proposed two models trained on 41,830 samples. The first model uses the ECG signal data, while the second uses ECG plot images. For each model, CNN layers were used with batch normalization and dropout for feature extraction, then a fully connected layer and SoftMax activation layer were used to classify eight heart abnormalities. Nugent et al. [24] presented a method based on sub-dividing the classification space into bi-group classifiers generated through the deployment of neural networks. An evidential reasoning framework was combined with this method to accommodate any conflicts among the bi-group classifiers. This method was able to classify six ECG signal types using 12-lead data.

The remainder of this paper is organized as follows: section 2 summarizes significant research on ECG classification, while section 3 details the method used in this study. Next, section 4 provides experimental details section 5 describes the findings. Finally, in section 6, conclusions are presented.

2.   METHOD AND MATERIALS
2.1. Dataset
The PhysioNet/Cinc 2020 challenge [25] dataset is used in this work. It is a publicly available, multi-class, and multi-labeled ECG signal dataset containing 43,101 labeled ECG records. Furthermore, the dataset was collected from different sources, as shown in Table 1, which makes it perfect for ensuring the suggested method's generalization ability.

Table 1. The dataset sources and the number of records for each sub-dataset

Dataset	# Records
The China Physiological Signal Challenge (CPSC) 2018 [26]	10,330
The St. Petersburg Institute of Cardiological Technics (INCART) database of 12-lead arrhythmias [27]	74
The PhysikalischTechnischeBundesanstalt (PTB) [28] and the more recent PTB-XL [29]	22,353
The Georgia 12-lead ECG Challenge (G12EC) database [25]	10,344

No distinction is established between the data sources in this experiment and all records are pooled into a single repository. Furthermore, the metadata of each record includes the individual's biological information and gender. The average age of the participants is 60 years old. Females account for 46.9% of the participants, while 53.1% are males. Figure 1 shows an ECG sample from the dataset. Most of the records contain more than one diagnosis, and the total number of unique combinations of diagnoses is 1,414. The details of the 27 ECG abnormalities are shown in Figure 2, from which we infer a significant dataset class imbalance. Sinus rhythm (NSR) is present in more than 20,000 recordings, whereas PVCs were detected in less than 200 samples. Such an imbalance could undermine the model performance, as the model is likely to learn the diagnostic pattern from categories with many samples while ignoring the minority categories.

Figure 1. Random ECG from the dataset

2.2. Data preparation

To avoid overfitting and test the model performance efficiently, the data is split into 34,480 samples for training and 8,621 samples for testing. Since each record in the dataset consists of 12-lead ECG sequences represented as multivariate time series, the correlation between the sequence ordering and the leads should be investigated to enhance model performance. While the samples came from different data sources, as mentioned earlier, truncating and padding were used to unify the sample lengths, and all the input records were fixed at 5,000 data points by padding the signals shorter than 5,000 data points with zeros. The input size for the suggested model is a matrix of size 5,000×12.

2.3. Problem statement

For 27 classes of 12-lead ECG signals, a pattern classification can be formulated to do the classifying. Each signal sample can be presented as a matrix of, so given sequence $X = \{x[0], x[1], x[2], \dots x[n]\}$, a classifier is trained to learn the class as in (1):

$$\hat{Y} = argmax\big(f(C = c|X)\big)c = 1, 2, \dots, 27 \qquad (1)$$

where C represents the labels for the record list, $x[n] \in \mathbb{R}^{27x1}$ is the input matrix for sample n, and $\hat{Y}$ is the class prediction.
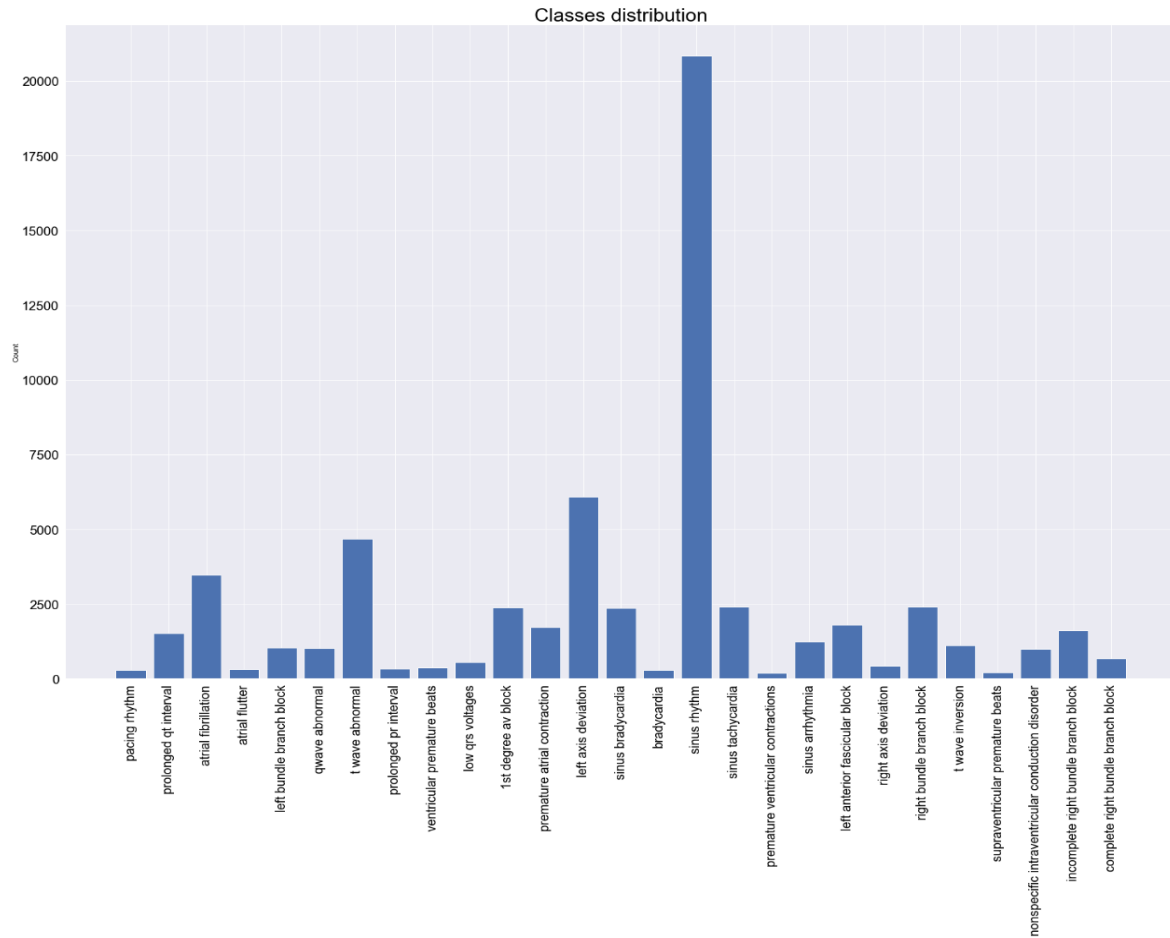
Figure 2. Total counts for 27 ECG abnormalities in the original dataset

3.   THE PROPOSED MODELS

This study adopted five DL models to achieve high classification performance. The prediction time is considered, so all suggested models (except the Inception-based model) are designed with as few layers and parameters as possible. The implementations of these proposed models are discussed next.

3.1. LSTM model

A DL model based on LSTM is implemented to achieve a high-recognition performance on ECG signals derived from 12 leads. There are three gates in the LSTM unit: the input, output, and forgetting gates (i, y, and f, respectively). In (2)-(4) are used to calculate the outputs of these gates, while c and h in (5) and (7) represent the cell state and the hidden state, respectively.

$$y_t = \tanh{(W_y x_t + R_y h_{t-1} + b_y)} \tag{2}$$

$$i_t = \sigma{(W_i x_t + R_i h_{t-1} + b_i + W_i \odot c_{t-1})} \tag{3}$$

$$f_t = \sigma{(W_f x_t + R_f h_{t-1} + b_f + W_f \odot c_{t-1})} \tag{4}$$

$$c_t = i_t \odot y_t + f_t \odot c_{t-1} \tag{5}$$

$$o_t = \sigma{(W_o x_t + R_o h_{t-1} + b_o + W_o \odot c_t)} \tag{6}$$

$$h_t = o_t \odot \tanh(c_t) \tag{7}$$

Where x_t is the input at time t;W, b, and R are the input weight, bias, and recurrent weight matrices of the LSTM unit, respectively;σ is the sigmoid function (σ(x)= 1/(1+e^(-x))); and ⊙ represents point-wise multiplication.

Two LSTM layers with 64 units each and one dense layer with 32 units (using Relu as the activation function) are constructed to build a simple optimized model. For a review of this type of model, refer to [30]. Bidirectional LSTM is used, so the signal streams forward and backward at each time step, and the outputs of both streams are combined to compute the temporal relationship. The dropout technique (with a 30% rate) is used during the training process to avoid overfitting. The dropout mechanism turns off 30% of the dense layer neurons during training, which regularizes the network. Furthermore, the model is trained just for ten epochs to avoid making the model memorize the training data. Figure 3 and Table 2 provide details of the layers' structure and the parameters of each layer to enable the reader to rebuild the proposed model.
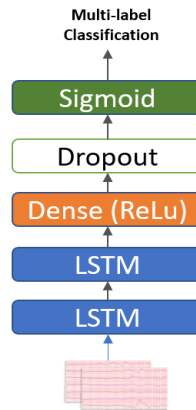

Figure 3. LSTM model's layer structure

Table 2. LSTM model's parameters

Layer	Details	# Parameters
LSTM_1	64 units	19,712
LSTM_2	64 units	33,024
Dense_1	32 units, Relu	2,080
Dropout	0.3	0
Dense_2	27 units, Sigmoid	759
Total # parameters	55,575	

3.2. The hybrid CNN-LSTM model

The CNN-LSTM design is used to handle the nonlinearity and complexity of the 12-lead data. In addition, the CNN-LSTM design is used for specific sequence prediction problems with spatial inputs, such as those related to video or audio [31]. Using this method, the CNN can learn the relevant features from the ECG signals coming from different leads, while the LSTM bridges a long-time lag between the inputs over arbitrary time intervals. Practical features can be obtained because the LSTM can depict temporal patterns at different frequencies. One 1-D CNN layer (with 64 filters and a kernel size of 8) and one max-pooling layer (used for dimensionality reduction and speeding up the training process) are followed by one LSTM layer with 128 units and a dense layer with 27 neurons (with a Sigmoid activation function for multi-label class prediction). Figure 4 and Table 3 provide the details of the layers' structure and the parameters of each layer to enable the reader to rebuild the proposed model.

3.3. The hybrid LSTM-GRU model

Cho *et al*. [32] introduced the gated recurrent unit (GRU), which can gather associations across timescales in an adaptive manner. As with the LSTM, each GRU employs gating units to control the flow of information inside the unit. A GRU is a simplified version of the LSTM, as it has only two gates (the reset r_t and update z_t gates in (8) and (9)) in its architecture. It offers outstanding performance and solves the vanishing gradient problem [33]. The hidden state at time t can be calculated via (10).

$$r_t = \sigma(W_r x_t + R_r h_{t-1}) \tag{8}$$

$$z_t = \sigma(W_z x_t + R_z h_{t-1}) \tag{9}$$

$$h_t = f(W x_t + h_{t-1}) \tag{10}$$

The update value of the unit activation at time t is z_t, which can be found using (11):

$$z_t = \sigma(W_z x_t + R_z h_{t-1}) \tag{11}$$

The candidate activation h̃_t is calculated as in (12):

$$\tilde{h}_t = \tanh(W x_t + R(r_t \oplus h_{t-1})) \tag{12}$$

where $\oplus$ is the element-wise multiplication process.

Combining the LSTM and GRU enables the model to learn the features of the time-space data for the ECG signals. The model consists of a 16-unit LSTM layer, a 100-unit GRU layer, a 32-neuron dense layer (with a Relu activation function), and a 27-neuron dense layer (with a Sigmoid activation function) for multi-label classification. The model architecture is explained in Figure 5, while Table 4 lists the layer parameters.

3.4. The hybrid CNN-GRU model

Similar to the hybrid CNN-LSTM model, this next model uses a GRU layer instead of the LSTM. The CNN can extract features perfectly, but as a feed-forward neural network, it does not have input memory and cannot cycle formed connections in time-based data. The GRU units with their gates can solve this issue and the vanishing gradient problem. This model features a 1-D CNN layer with max-pooling for dimensionality reduction, followed by the GRU layer. The detailed model architecture is shown in Figure 6 and Table 5.

3.5. The Inception-ResNet-v2 model

In this model, the Inception-ResNet-v2 [17] network is utilized to categorize ECG signals. The network's architecture is depicted in Figure 7. It comprises three parts. In the first part, the stem has nine convolutional layers, and two max-pooling layers are used to pre-process the original input before it enters the Inception-ResNet blocks. The second part is illustrated in Figure 8. Figure 8(a) illustrates the Inception-ResNet-A with two 3×3 inception kernels. Figure 8(b) for dimentionalty improvement, and Figure 8(c) illustrates the Inception-ResNet-B with an asymmetric filter combination of one 1x7 filter and one 7×1 filter in the inception module. Figure 8(e) illustrates the Inception-ResNet-C with a small and asymmetric filter combination of one 1×3 filter and one 3×1 filter; 1×1 convolutions are utilized prior to the large filters in these blocks. Through asymmetric convolution splitting, the network increases the diversity of the filter patterns. In addition, the reductions in Figures A and C shown in Figure 8(d) are performed to enhance the dimension, which must balance for the Inception block's dimensionality reduction. The final part is the classification layer, which includes the pooling and Sigmoid algorithm.
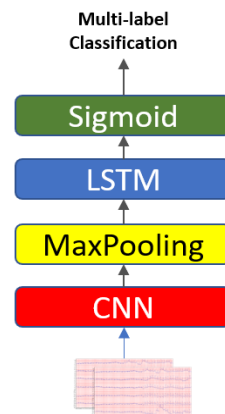

Figure 4. CNN-LSTM model's layer structure

Table 3. CNN-LSTM model's parameters

Layer	Details	# Parameters
CNN_1	64 units, kernel_size=8, strides=1, Relu	6.208
MaxPooling	Pool_size=4	0
LSTM_1	128 units	98,816
Dense_1	27 units, Sigmoid	2,967
Total # parameters	107,991	
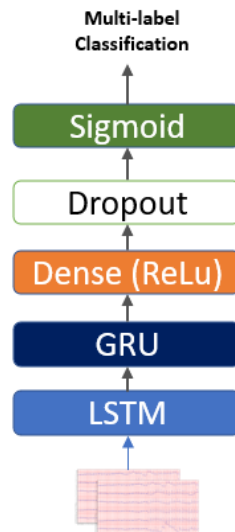
Figure 5. LSTM-GRU model's layer structure

Table 4. LSTM-GRU model's parameters

Layer	Details	# Parameters
LSTM_1	16 units	1,856
GRU	100 units	35,100
Dense_1	32 units, Relu	3,232
Dropout	0.3	0
Dense_2	27 units, Sigmoid	759
Total # parameters	40,947	
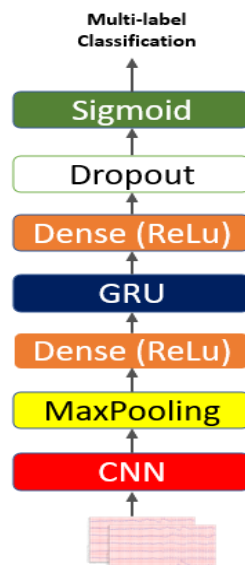
Figure 6. CNN-GRU model's layer structure

Table 5. CNN-GRU model's parameters

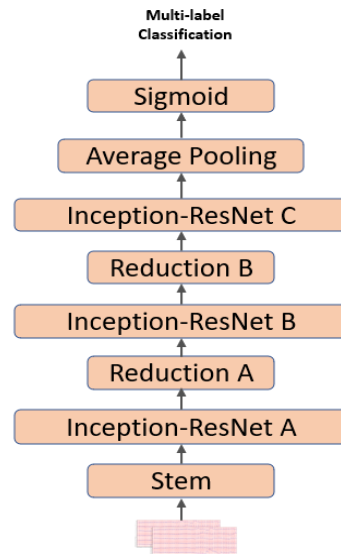| Layer	Details	# Parameters
CNN	64 units, kernel_size=8, strides=1, Relu	592
MaxPolling	Pool_size=2	0
Dense_1	32 units, Relu	544
GRU	100 units	39,900
Dense_2	32 units, Relu	3,232
Dropout	0.3	0
Dense_3	27 units, Sigmoid	759
Total # parameters		45,027

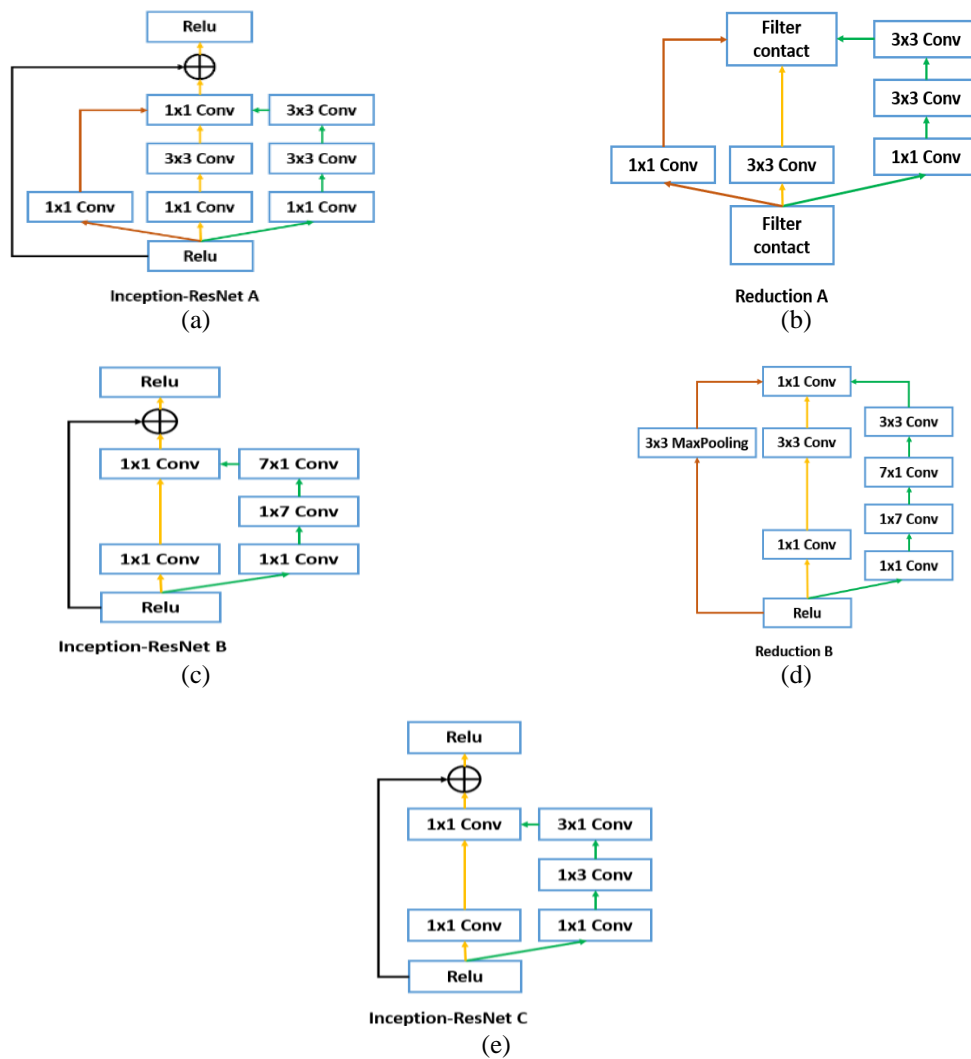Figure 7. The architecture of the Inception-ResNet

Figure 8. The architecture of the Inception-ResNet (a) inception-ResNet A, (b) reduction A, (c) inception-ResNet B, (d) reduction B, and (e) inception-ResNet C

4. RESULTS AND DISCUSSION
4.1. Computing environment

The training was conducted on a 1.7 GHz Intel Core i5 processor, 8 GB of RAM, a 64-bit instruction set on Windows 10 Pro, and a display card with a memory capacity of 2 GB from NVIDIA. Python was utilized as the primary programming language, and the TensorFlow package was leveraged to build the CNN model.

4.2. Loss function

In this work, the binary cross-entropy loss function is used. There are many classes in most samples, and it calculates the probability of each class in each sample. In (13) shows how the binary cross-entropy loss is calculated:

$$L_{BCE} = -\frac{1}{N}\sum_i^N \sum_j^M (p(x_{ij}).\log q(x_{ij}) + \left(1 - p(x_{ij}).log\left(1 - q(x_{ij})\right)\right) \tag{13}$$

Where:
p(x) is the probability of class x in the target
q(x) is the probability of class x in the prediction
N is the number of samples, and M is the number of classes

4.3. Experimental setup

The parameters of all the models examined in this study are initialized randomly, and for optimization, the Adam optimization function is used with a 0.001 initial learning rate. For callbacks, reduced learning is used with one patience so that the learning rate will be reduced by a factor of 0.1 when there is no improvement in the loss of the validation data after one epoch. Furthermore, to avoid overfitting, the early stopping technique is used with patience of 2 to curtail training when there is no improvement in the validation data loss after two epochs. Table 6 shows the training time for one epoch for each model.

Table 6. The training time (in seconds) for one epoch for each model

Model	The training time for one epoch (in sec)
LSTM	1,241
CNN-LSTM	153
GRU-LSTM	9,625
CNN-GRU	3,814
Inception	985

4.4. Evaluation metrics

For evaluating the performance of these methods, four commonly used performance metrics are used in this study, namely, accuracy (14), recall (15), and precision (16), and the area under the curve (AUC) [34].

$$A_{cc} = \frac{True_p + True_n}{True_p + False_p + True_n + False_n} \tag{14}$$

$$R_{call} = \frac{True_p}{True_p + False_n} \tag{15}$$

$$Pre = \frac{True_p}{True_p + False_p} \tag{16}$$

Here, $True_p$ denotes a true positive, $True_n$ denotes a true negative, $False_p$ indicates a false positive, and $False_n$ indicates a false negative. The AUC is the performance aggregation measure across all e possible thresholds for classification. The AUC value is contained in [0,1], and a higher value means better model performance.

5. RESULTS

As mentioned earlier, the dataset is split into 34,480 samples for training and 8,621 samples for testing. The loss function graphs for all the suggested models are shown in Figure 9 (in Appendix). Note the

stability of all models (convergence) and the low variability between the training and validation data. However, except for the CNN-LSTM model, the loss graphs indicate some instability and high variability, a sign of overfitting. Figure 10 (in Appendix) shows the accuracy and precision graphs for the training and validation data over the entire dataset; note that the models mostly converge after six to eight epochs.

The performances of all the proposed models are measured using the test data (8,621 samples that the models have not trained on before). Table 7 shows the performance of each model across the entire dataset. The results are further analyzed in the discussion section.

Table 7. The performance metrics for the proposed models across the entire dataset

Model	Accuracy	Recall	Precision	AUC	Loss
LSTM	0.96	0.27	0.71	0.62	0.15
CNN-LSTM	0.95	0.17	0.58	0.53	0.16
LSTM-GRU	0.96	0.30	0.80	0.70	0.14
CNN-GRU	0.96	0.30	0.82	0.67	0.14
Inception	0.97	0.38	0.84	0.77	0.11

6. DISCUSSION

This study is one of the first studies to use DL to diagnose 27 cardiac abnormalities automatically based on a large volume of data on 12-lead ECGs. We have shown that a DL technique is capable of accurately categorizing 12-lead ECG results. Additionally, the Inception model had a high accuracy of 0.97, while the other models had accuracies of 0.96. These findings suggest that a DL method will be useful for ECG triage and able to minimize the clinical workload through enhanced prioritizing of ECGs for interpretation by a cardiologist. Note that non-cardiologists accurately diagnose 35% to 95% of cardiac issues, with considerable variance among physicians and increases in performance with experience [35]–[37].

The dataset used in this study is challenging because it came from different sources and the classes are imbalanced. Furthermore, solving multi-label classification problems is more complicated than solving multi-class problems [38]. Nevertheless, Table 7 shows the high precision-low recall obtained results for all the models, indicating that when it was difficult to label a sample, the models chose not to predict an incorrect label, increasing the false-negative error. According to Figures 11 and 12 (in Appendix), most models converged after a few epochs (six to eight epochs in most cases). The Inception model, which uses residual networks and CNN, obtained the best performance.

In addition, the other models achieved close results despite their simple structures. Of the models that used CNN for feature extraction (CNN-LSTM and CNN-GRU), the results obtained by the GRU-based model were significantly better in terms of precision, recall, and the AUC. On the other hand, the GRU-LSTM obtained results close to those of the CNN-GRU model but with a higher training time (see Table 6), as the max-pooling layer in the CNN-GRU model provides dimensionality reduction. The diversity of the dataset's sources is an advantage for testing the generalization abilities of the models. Furthermore, we trained the models using the PTB-XL dataset to test the proposed models on a single-source dataset. The results in Table 8 and Figures 11 and 12 (in Appendix) show significant increases in recall value (lower false negative) and the AUC.

Table 8. The performance metrics for the proposed models on the PTB-XL dataset

Model	Accuracy	Recall	Precision	AUC	Loss
LSTM	0.95	0.48	0.83	0.59	0.16
CNN-LSTM	0.95	0.48	0.83	0.55	0.16
LSTM-GRU	0.95	0.49	0.84	0.69	0.15
CNN-GRU	0.95	0.49	0.84	0.63	0.15
Inception	0.97	0.64	0.87	0.84	0.10

6.1. Comparison with other models

Many methods have been suggested in the literature for ECG signal classification. Nevertheless, the number of classes and leads used differ across these studies, which should be considered when comparing their performances. Table 9 compares this paper's methods and some related approaches (N and MI stand for normal class and myocardial infarction, respectively). The comparison shows that our Inception method outperformed all other methods, although it classifies 27 multi-label classes and uses a dataset based on different sources. Total counts for 27 ECG abnormalities in the original dataset and their corresponding abbreviations are explained in Table 10.

Table 9. Comparison of the methods with related methods

Study	# Leads	# Classes	Dataset	Method	Accuracy
[39]	12	2	N 31,722 MI 49,930	CNN	0.935
[40]	3	2	N 5,000 MI 15,000	Fourier/logistic regression	0.956
[41]	12	3	PTB and AF-challenge	CNN-LSTM	0.946
[16]	12	7	7,704 samples	ResNet-LSTM	0.81
This study	12	27	[25]	Inception	0.97

Table 10. ECG abnormalities and their corresponding abbreviations

ECG abnormality	Abbreviation
1st degree AV block	IAVB
Atrial fibrillation	AF
Atrial flutter	AFL
Bradycardia	Brady
Complete right bundle branch block	CRBBB
Incomplete right bundle branch block	IRBBB
Left anterior fascicular block	LAnFB
Left axis deviation	LAD
Left bundle branch block	LBBB
Low QRS voltages	LQRSV
Nonspecific intraventricular conduction disorder	NSIVCB
Pacing rhythm	PR
Premature atrial contraction	PAC
Premature ventricular contractions	PVC
Prolonged PR interval	LPR
Prolonged QT interval	LQT
Q wave abnormal	QAb
Right axis deviation	RAD
Right bundle branch block	RBBB
Sinus arrhythmia	SA
Sinus bradycardia	SB
Sinus rhythm	NSR
Sinus tachycardia	STach
Supraventricular premature beats	SVPB
T wave abnormal	Tab
T wave inversion	TInv
Ventricular premature beats	VPB

7. CONCLUSION

This work presents an end-to-end method for automatic 12-lead ECG classification. Among five suggested models, the Inception network-based model achieved the best performance, with an accuracy of 0.97. The suggested model classifies 27 multi-label abnormalities indicated by 12-lead ECG signals, while the related methods classify nine types, at most. Experiments show that our suggested model outperforms the other related models on a large dataset based on different sources and a single dataset from the same source. Additionally, because all of the datasets utilized are real-world data, we feel that this approach can be developed and applied in the medical field or used as a screening tool in conditions/locations where access to a 12-lead ECG is limited. We suggest solving the imbalanced dataset problem in future works, which can be done by collecting more samples or through various techniques, such as data generation, to improve the classification of rare heart abnormalities.
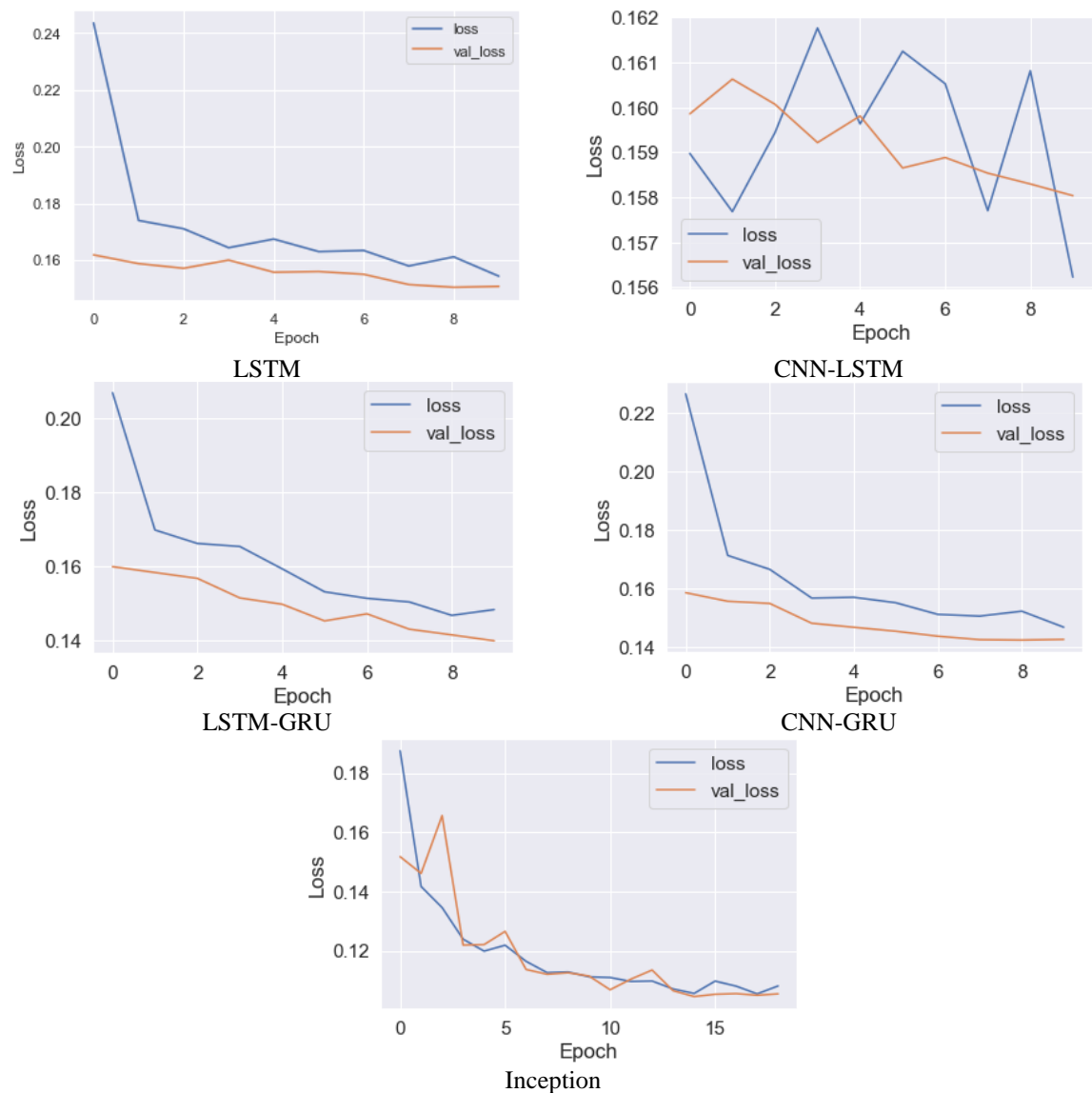
APPENDIX

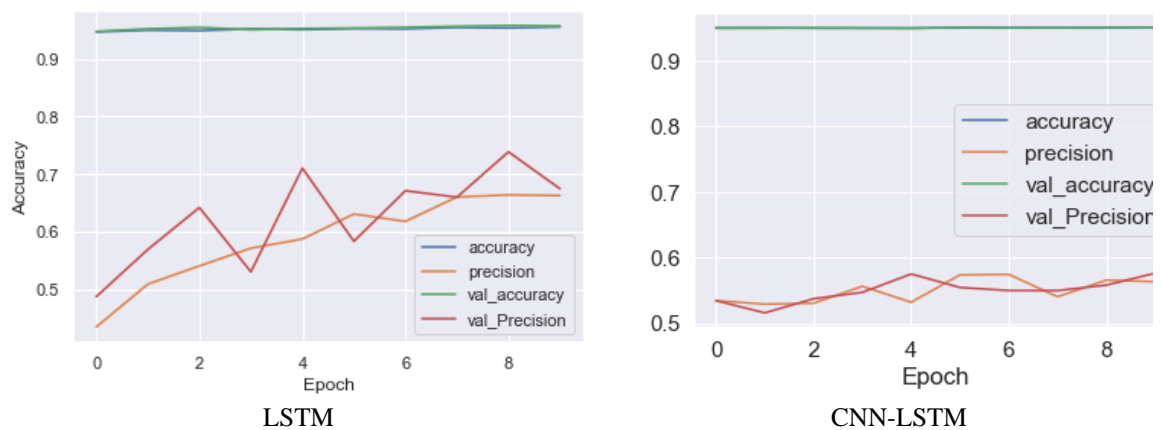Figure 9. Loss per epoch for the models trained on the entire dataset

Figure 10. Accuracy per epoch for the models trained on the entire dataset
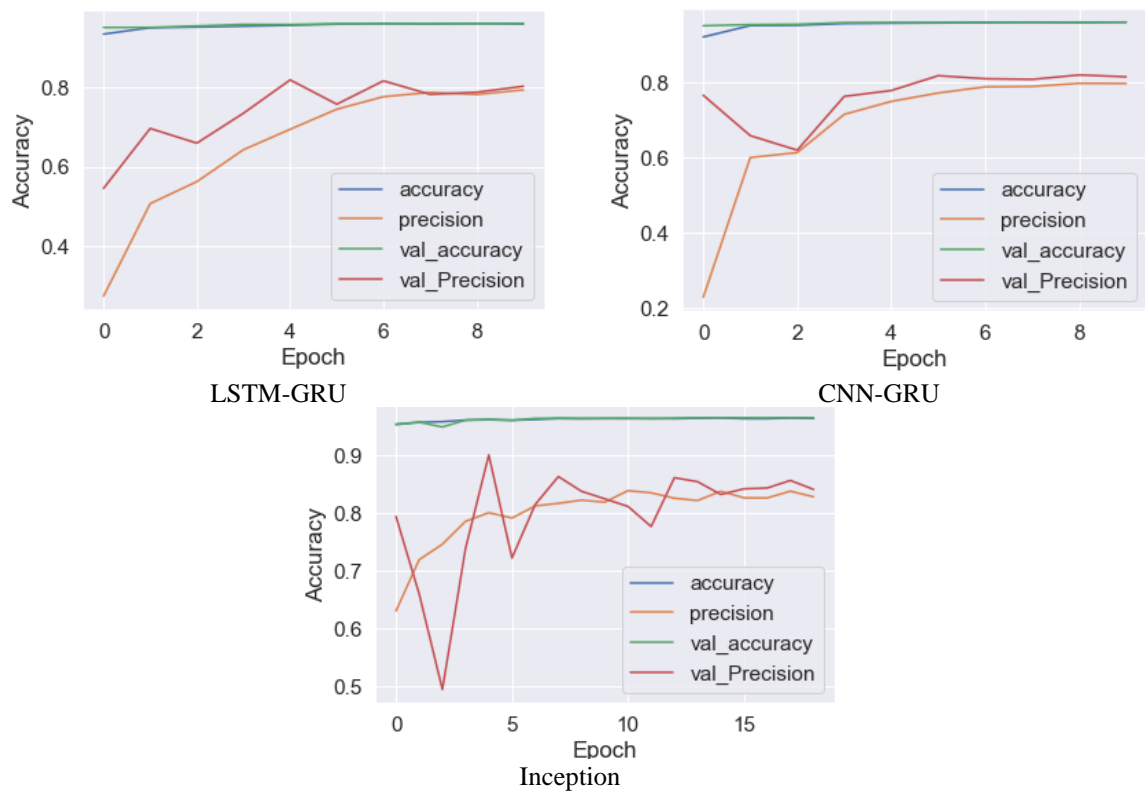
Figure 10. Accuracy per epoch for the models trained on the entire dataset (continue)
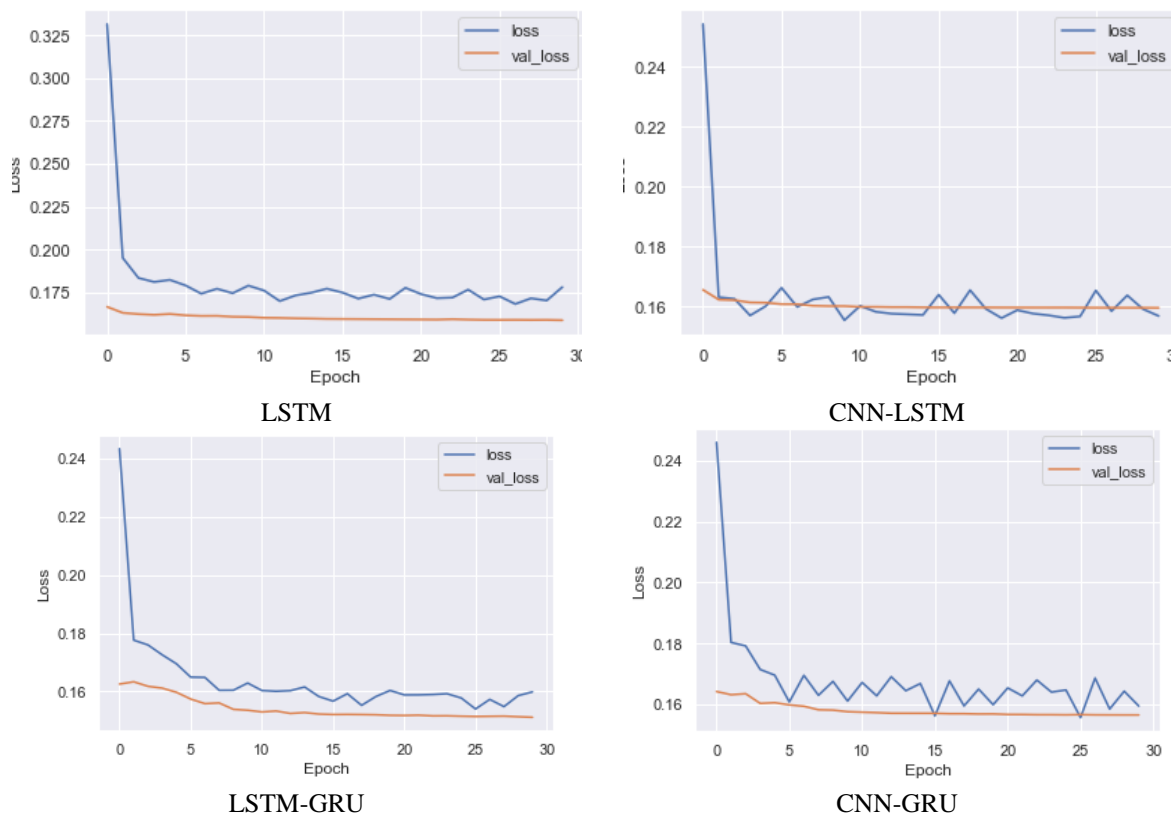

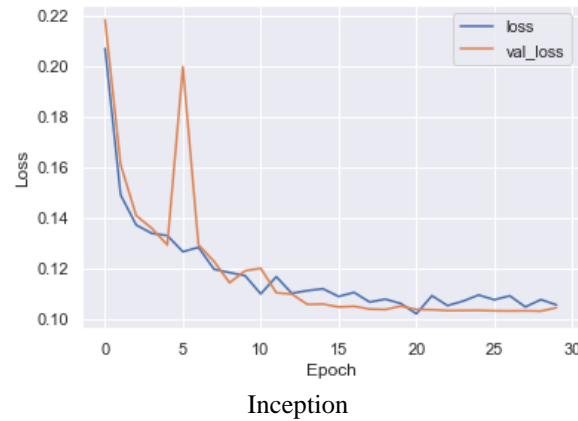Figure 11. Loss per epoch for the models trained on the PTB-XL dataset

Inception

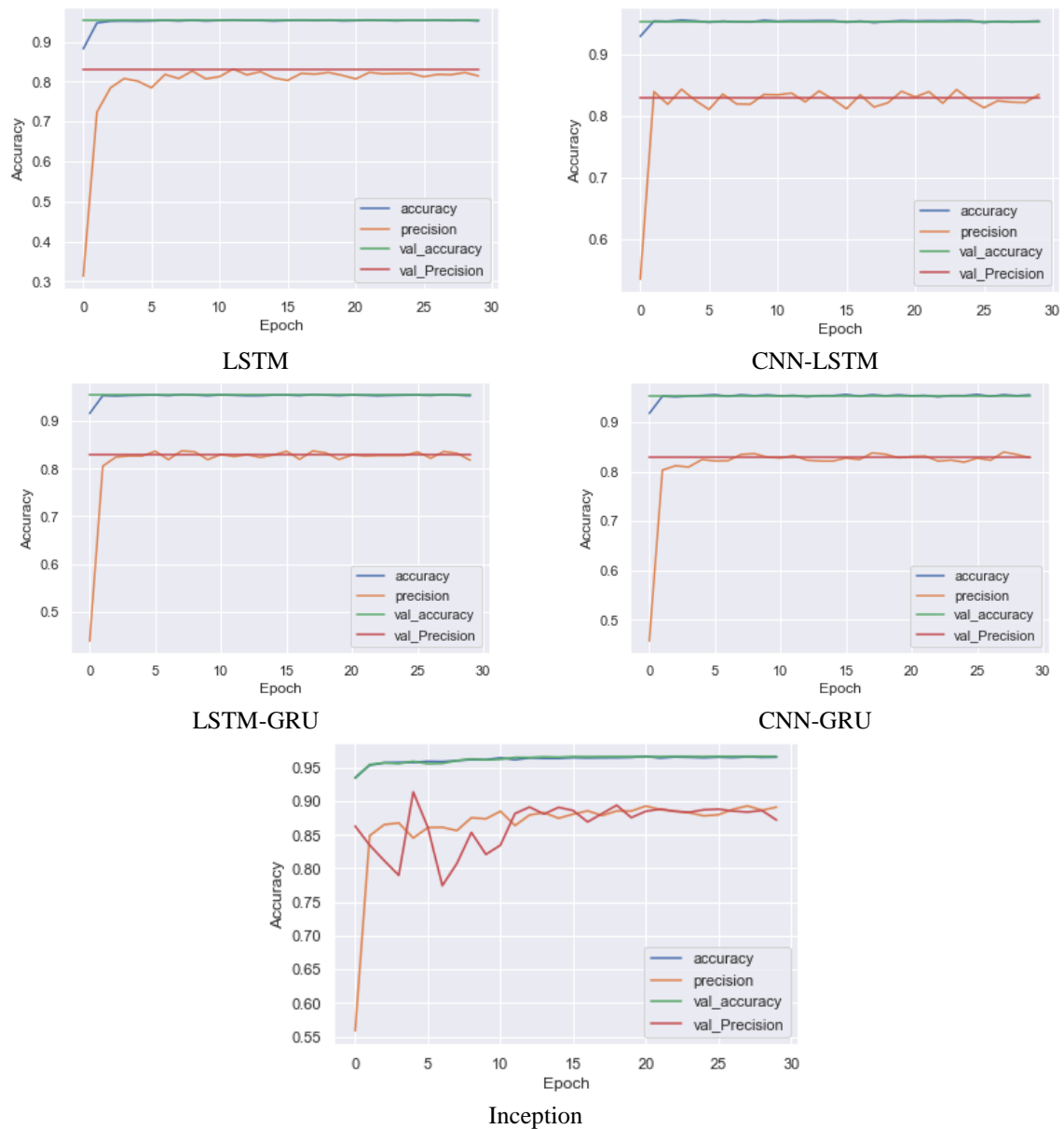Figure 11. Loss per epoch for the models trained on the PTB-XL dataset (continue)

LSTM   CNN-LSTM

LSTM-GRU   CNN-GRU

Inception

Figure 12. Accuracy per epoch for the models trained on the PTB-XL dataset

REFERENCES

[1]   J. R. Cox, F. M. Nolle, and R. M. Arthur, "Digital analysis of the electroencephalogram, the blood pressure wave, and the electrocardiogram," *Proceedings of the IEEE*, vol. 60, no. 10, pp. 1137–1164, 1972, doi: 10.1109/proc.1972.8877.

[2]   F. Amato, A. López, E. M. P-Méndez, P. Vaňhara, A. Hampl, and J. Havel, "Artificial neural networks in medical diagnosis," *Journal of Applied Biomedicine*, vol. 11, no. 2, pp. 47–58, Jul. 2013, doi: 10.2478/v10136-012-0031-x.

[3]   M. Bickerton and A. Pooler, "Misplaced ECG electrodes and the need for continuing training," *British Journal of Cardiac Nursing*, vol. 14, no. 3, pp. 123–132, Mar. 2019, doi: 10.12968/bjca.2019.14.3.123.

[4]   D. Eslava, S. Dhillon, J. Berger, P. Homel, and S. Bergmann, "Interpretation of electrocardiograms by first-year residents: the need for change," *Journal of Electrocardiology*, vol. 42, no. 6, pp. 693–697, Nov. 2009, doi: 10.1016/j.jelectrocard.2009.07.020.

[5]   C. Alexakis *et al.*, "Feature extraction and classification of electrocardiogram (ECG) signals related to hypoglycaemia," in *Computers in Cardiology, 2003*, 2003, doi: 10.1109/cic.2003.1291211.

[6]   S. Mahmoodabadi, A. Ahmadian, and M. Abolhasani, "ECG feature extraction using Daubechies wavelets," Jan. 2005, pp. 343–348.

[7]   J. P. Martinez, R. Almeida, S. Olmos, A. P. Rocha, and P. Laguna, "A Wavelet-Based ECG Delineator: Evaluation on Standard Databases," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 4, pp. 570–581, Apr. 2004, doi: 10.1109/tbme.2003.821031.

[8]   R. R. van de Leur *et al.*, "Automatic Triage of 12-Lead ECGs Using Deep Convolutional Neural Networks," *Journal of the American Heart Association*, vol. 9, no. 10, May 2020, doi: 10.1161/jaha.119.015138.

[9]   J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." arXiv, May 24, 2019, doi: 10.48550/arXiv.1810.04805.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017, doi: 10.1145/3065386.

[11] Z. Xiong, M. P. Nash, E. Cheng, V. V. Fedorov, M. K. Stiles, and J. Zhao, "ECG signal classification for the detection of cardiac arrhythmias using a convolutional recurrent neural network," *Physiological Measurement*, vol. 39, no. 9, p. 094006, Sep. 2018, doi: 10.1088/1361-6579/aad9ed.

[12] P. Sodmann, M. Vollmer, N. Nath, and L. Kaderali, "A convolutional neural network for ECG annotation as the basis for classification of cardiac rhythms," *Physiological Measurement*, vol. 39, no. 10, p. 104005, Oct. 2018, doi: 10.1088/1361-6579/aae304.

[13] A. Elola *et al.*, "Deep Neural Networks for ECG-Based Pulse Detection during Out-of-Hospital Cardiac Arrest," *Entropy*, vol. 21, no. 3, p. 305, Mar. 2019, doi: 10.3390/e21030305.

[14] P. A. Warrick and M. N. Homsi, "Ensembling convolutional and long short-term memory networks for electrocardiogram arrhythmia detection," *Physiological Measurement*, vol. 39, no. 11, p. 114002, Oct. 2018, doi: 10.1088/1361-6579/aad386.

[15] S. Hong, C. Xiao, T. Ma, H. Li, and J. Sun, "MINA: Multilevel Knowledge-Guided Attention for Modeling Electrocardiography Signals," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, Aug. 2019, doi: 10.24963/ijcai.2019/816.

[16] Y.-J. Chen, C.-L. Liu, V. S. Tseng, Y.-F. Hu, and S.-A. Chen, "Large-scale Classification of 12-lead ECG with Deep Learning," in *2019 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, May 2019, doi: 10.1109/bhi.2019.8834468.

[17] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, Feb. 2017, doi: 10.1609/aaai.v31i1.11231.

[18] Z. Liu, B. Chen, and A. Zhang, "Building segmentation from satellite imagery using U-Net with ResNet encoder," in *2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE)*, Dec. 2020, doi: 10.1109/icmcce51767.2020.00431.

[19] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-Excitation Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020, doi: 10.1109/tpami.2019.2913372.

[20] J. Pan and W. J. Tompkins, "A Real-Time QRS Detection Algorithm," *IEEE Transactions on Biomedical Engineering*, vol. BME-32, no. 3, pp. 230–236, Mar. 1985, doi: 10.1109/tbme.1985.325532.

[21] U. B. Baloglu, M. Talo, O. Yildirim, R. S. Tan, and U. R. Acharya, "Classification of myocardial infarction with multi-lead ECG signals and deep CNN," *Pattern Recognition Letters*, vol. 122, pp. 23–30, May 2019, doi: 10.1016/j.patrec.2019.02.016.

[22] N. Fayyazifar, S. Ahderom, D. Suter, A. Maiorana, and G. dwivedi, "Impact of Neural Architecture Design on Cardiac Abnormality Classification Using 12-lead ECG Signals," in *2020 Computing in Cardiology Conference (CinC)*, Dec. 2020, doi: 10.22489/cinc.2020.161.

[23] V. Gliner, N. Keidar, V. Makarov, A. I. Avetisyan, A. Schuster, and Y. Yaniv, "Automatic classification of healthy and disease conditions from images or digital standard 12-lead electrocardiograms," *Scientific Reports*, vol. 10, no. 1, Oct. 2020, doi: 10.1038/s41598-020-73060-w.

[24] C. D. Nugent, J. A. C. Webb, N. D. Black, G. T. H. Wright, and M. McIntyre, "An intelligent framework for the classification of the 12-lead ECG," *Artificial Intelligence in Medicine*, vol. 16, no. 3, pp. 205–222, Jul. 1999, doi: 10.1016/s0933-3657(99)00006-8.

[25] E. A. P. Alday *et al.*, "Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020," *Physiological Measurement*, vol. 41, no. 12, p. 124003, Dec. 2020, doi: 10.1088/1361-6579/abc960.

[26] F. Liu *et al.*, "An Open Access Database for Evaluating the Algorithms of Electrocardiogram Rhythm and Morphology Abnormality Detection," *Journal of Medical Imaging and Health Informatics*, vol. 8, no. 7, pp. 1368–1373, Sep. 2018, doi: 10.1166/jmihi.2018.2442.

[27] E. Yakushenko, "St Petersburg INCART 12-lead Arrhythmia Database," *PhysioBank, PhysioToolkit, and PhysioNet*, May 2008, doi: 10.13026/C2V88N.

[28] R. Bousseljot, D. Kreiseler, and A. Schnabel, "Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet," *Biomedizinische Technik/Biomedical Engineering*, pp. 317–318, Jul. 2009, doi: 10.1515/bmte.1995.40.s1.317.

[29] P. Wagner *et al.*, "PTB-XL, a large publicly available electrocardiography dataset," *Scientific Data*, vol. 7, no. 1, May 2020, doi: 10.1038/s41597-020-0495-6.

[30] K. Greff, R. K. Srivastava, J. Koutnik, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A Search Space Odyssey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 10, pp. 2222–2232, Oct. 2017, doi: 10.1109/tnnls.2016.2582924.

[31] T.-Y. Kim and S.-B. Cho, "Web traffic anomaly detection using C-LSTM neural networks," *Expert Systems with Applications*, vol. 106, pp. 66–76, Sep. 2018, doi: 10.1016/j.eswa.2018.04.004.

[32] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, "On the Properties of Neural Machine Translation: Encoder–Decoder Approaches," in *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, 2014, doi: 10.3115/v1/w14-4012.

[33] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling." arXiv, Dec. 11, 2014, doi: 10.48550/arXiv.1412.3555.

[34] W. Zhu, N. Zeng, and N. Wang, "Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS ® Implementations," *NorthEast SAS users group, health care and life sciences*, Jan. 2010.

[35] G. Veronese *et al.*, "Emergency physician accuracy in interpreting electrocardiograms with potential ST-segment elevation myocardial infarction: Is it enough?," *Acute Cardiac Care*, vol. 18, no. 1, pp. 7–10, Jan. 2016, doi: 10.1080/17482941.2016.1234058.

[36] J. J. Goy, J. Schlaepfer, and J. C. Stauffer, "Competency in interpretation of 12-lead electrocardiogram among Swiss doctors," *Swiss Medical Weekly*, May 2013, doi: 10.4414/smw.2013.13806.

[37] S. M. Salerno, P. C. Alguire, and H. S. Waxman, "Competency in Interpretation of 12-Lead Electrocardiograms: A Summary and Appraisal of Published Evidence," *Annals of Internal Medicine*, vol. 138, no. 9, p. 751, May 2003, doi: 10.7326/0003-4819-138-9-200305060-00013.

[38] K. Dembczyński, W. Waegeman, W. Cheng, and E. Hüllermeier, "On label dependence and loss minimization in multi-label classification," *Machine Learning*, vol. 88, no. 1–2, pp. 5–45, Jun. 2012, doi: 10.1007/s10994-012-5285-8.

[39] A. M. Lodhi, A. N. Qureshi, U. Sharif, and Z. Ashiq, "A Novel Approach Using Voting from ECG Leads to Detect Myocardial Infarction," in *Advances in Intelligent Systems and Computing*, Springer International Publishing, 2018, pp. 337–352, doi: 10.1007/978-3-030-01057-7_27.

[40] D. Sadhukhan, S. Pal, and M. Mitra, "Automated Identification of Myocardial Infarction Using Harmonic Phase Distribution Pattern of ECG Data," *IEEE Transactions on Instrumentation and Measurement*, vol. 67, no. 10, pp. 2303–2313, Oct. 2018, doi: 10.1109/tim.2018.2816458.

[41] H. W. Lui and K. L. Chow, "Multiclass classification of myocardial infarction with convolutional and recurrent neural networks for portable ECG devices," *Informatics in Medicine Unlocked*, vol. 13, pp. 26–33, 2018, doi: 10.1016/j.imu.2018.08.002.

BIOGRAPHIES OF AUTHORS

Atiaf A. Rawi 🆔 🇬 ⓢⓒ ℂ is now a doctoral studentof philosophy in computer science, Gezira University, Sudan. She earned her BSc in2010 and MSc in 2013 in computer science from Al-Anbar University, Anbar, Iraq. In 2014, She has worked as a lecturer at Al-Ma'aref University in (Iraq) for four years. Through this period, she has taught three subjects, such as compiler, logic design, and image processing and she has supervised several of the B. Sc's projects in the image processing field. Her current research is image processing, medical images and signals, classification, machine learning, and deep learning. She can be contacted at email: atiaf.ayal88@gmail.com.

Murtada Khalafallah Elbashir 🆔 🇬 ⓢⓒ ℂ received the B.Sc. degree (Hons.) in computer/statistics from the University of Gezira, Wad Madani, Sudan, in 2000, the M.Sc. degree (Hons.) in computer information systems from the University of the Free State, Bloemfontein, South Africa, in 2003, and the Ph.D. degree in computer science and technology from Central South University, Changsha, China, in 2013. In 2016, he promoted to an Associate Professor with the University of Gezira. He is currently an Associate Professor with the College of Computer and Information Sciences, Jouf University, Sakaka, Saudi Arabia. His current research interests include machine learning and bioinformatics. He can be contacted at email: mkelfaki@ju.edu.sa.

Awadallah M. Ahmed 🆔 🇬 ⓢⓒ ℂ received his B.Sc. degree (Hons.) in mathematical and computer sciences (Computer/Statistics) from University of Gezira, Sudan in 2004. He obtained his M.Sc. in computer science and information from University of Gezira, Sudan in 2010. He received his Ph. D degree in computer science from Sudan University of Science and Technology, Sudan in 2016. He is currently an Assistant professor with the Faculty of Mathematical and Computer Sciences, University of Gezira, Wad madani, Sudan. His research interests include computer networks, combinatorial optimization, and machine learning. He can be contacted at email: awadallah@uofg.edu.sd.