Classification of Premature Ventricular Contraction Using Deep Learning

Fabiola De Marco[1,2], Dewar Finlay[2], Raymond R. Bond[2]

[1] University of Salerno, Salerno, Italy Institution, City, Country
[2] Ulster University, Jordanstown, UK

Abstract

Electrocardiogram (ECG) analysis has been used to identify different heart problems and deep learning is emerging as a common tool to analyse ECGs. Premature ventricular contraction (PVC) is the most common cause of abnormal heartbeats; in most cases this is harmless but under specific conditions, it can lead to a life-threatening cardiac disease. Automated PVC detection in this scenario is a task of significant importance for relieving the heavy workloads of experts in the manual analysis of long-term ECGs. To identify PVCs, this research aims to use the MIT-BIH Arrhythmia Database to classify QRS complexes using five different deep neural networks: Long Short Term Memory, AlexNet, GoogleNet, Inception V3 and ResNet-50. The results showed high efficiency and reliability in the final diagnoses during two separate experiments (one with the entire dataset and the other with a balanced dataset). The ResNet-50 was the first experiment's best classifier (accuracy = 99.8%, F1-score = 99.2%), and the second experiment's best classifier was Inception V3 (accuracy = 98.8%, F1-score=98.8%). Relevant information, in this research, was extrapolated from a study of the confusion matrix to conduct a "failure analysis" to understand where and why the classifiers made incorrect classifications.

1. Introduction

Currently all over the world, particularly in countries with a predominantly western lifestyle, cardiovascular disease (CVD) is often the leading cause of death and despite the rapid growth of new technologies, the electrocardiogram (ECG) is still the main tool for analyzing and interpreting heart's rhythm and its electrical activity. A healthy person's heartbeat has four characteristics: P-wave, QRS complex, T-wave, U-wave and heart disease may be detected by interpreting these wave variations. Therefore this field of research will contribute not only to advancement in cardiology but also the enhancement of the patient's health.

Premature ventricular contraction (PVC) is a fairly common event that occurs in many people and causes additional heartbeats that begin in either of the heart's two ventricles and lead to QRS generation complexes of strange and large waveforms[1]. Usually, PVC is not harmful, but repeated PVCs may increase the risk of developing arrhythmias or cardiomyopathy [2] or, in the worst case, followed by other heart disorders may cause risky heart rhythms[3]. Automatic detection of PVC is an important challenge in the medical health domain since the conventional approach (medical or specialist workers who analyze and identify the ECG's features) is too sluggish and unreliable and because the presence of PVCs can be challenging for algorithms monitoring the heart rhythm in the decision-making process. Many studies in the literature have examined types of arrhythmias[4] and the issues of PVC-detection methods (limited dataset in ECG recording numbers, presence of noise or high inter and intra-patient variability), developing different models to overcome these problems by using the MIT-BIH Arrhythmia database as well. The situation in which ECG research is being performed today indicates a considerable increase in the use of deep learning techniques by developing models for the detection of PVC in children[5] that are revealed to be more efficient, models in which deep learning has been combined with rule inference [6] and also models used as the next step after the feature extraction phase [7]. Consequently, this project aims to use supervised learning to train five deep neural networks to classify QRS complexes in order to identify non-PVC records from PVC records. The research questions are: What is the performance of the different deep learning models for PVC detection? What type of deep learning model is the best classifier for detecting PVCs, such as CNNs or LSTM? What is the impact of pre-training on the performance of PVC detection deep learning models? What insights failure analysis provide into the false positive and false negative errors made by the classifier?

The remainder of the paper will present the methodology, the results and the discussion and the conclusion.

ISSN: 2325-887X DOI: 10.22489/CinC.2020.311

2. Methodology

In this research, five different deep neural networks, Long Short Term Memory (LSTM), AlexNet, GoogleNet, Inception V3 and ResNet-50, have been trained and the technique used is divided into two main strategies: *training with raw sample voltage data* and *training with images*. After training and testing, the results are compared with specific focus on the confusion matrix in order to do a ***failure analysis***. The general procedure is shown in the Fig.1

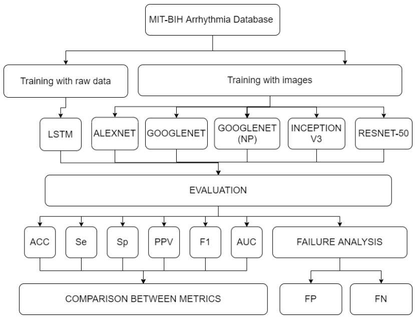


Figure 1: Activities Flow.

2.1. ECG database

The dataset used is the ”MIT-BIH Arrhythmia Database” developed by the laboratories at Boston’s Beth Israel Hospital (now the Beth Israel Deaconess Medical Center) and it is made up of ECGs from 47 patients of which the 60% from a mixed population of inpatients and the 40% from outpatients. For each patient, PVCs and QRS complexes were extracted, resulting in 7130 PVC records and 75048 non-PVC records, totalling 82,178 elements.

For each neural network we carried out two types of experiments were conducted:

- The first experiment uses the modified dataset, where only two labels are required to obtain a binary classification: labels 0 for non-PVC and labels 1 for PVC. But the dataset was not balanced in this situation since the number of non-PVCs was greater than the number of PVCs.
- The second experiment uses the same dataset but is balanced using class balancing techniques.

The patient’s information is not considered for the classification in both experiments because it could affect the results.

2.2. Training with raw sample data

The original dataset consists of raw sample data, in which each column refers to one sample of an ECG signal. For each annotation (R peak) provided from the database, we extracted 30 samples before this R peak and 30 samples after the R peak; hence each observation (row) represents 169 milliseconds. Only LSTM, a Recurrent Neural Network[8], was used for this strategy as it was specifically designed to prevent long-term dependencies in time series data problems.

2.3. Training with images

The raw sample data were converted into images to train Convolutional Neural Networks with transfer learning technique. These networks have the benefit of allowing the tacit presumption that the inputs are images, which enables the models to be more efficient by minimizing the number of network parameters. The Convolutional Neural Networks used were: AlexNet (AN), GoogleNet (GN), Inception V3 (I-V3), and ResNet-50 (RS-50); only the GoogleNet was used in both pre-trained and non pre-trained (NP) mode to better compare the non pre-trained LSMT.

3. Evaluation

Validating the classifier’s performance is an essential part of any project since the selection of metrics influences how the performance of machine learning algorithms is measured and compared. The performance of classifiers is measured by standard metrics: Accuracy(Acc), Sensitivity (Se), Specificity (Sp), Positive Predictivity (PPV) and also F1-score (F1) which is the harmonic mean between sensitivity and specificity and the Area Under the Curve (AUC).

4. Results

To validate the models used in this study, the dataset was split randomly, for each experiment, in training and testing for the LSTM, and training, validation and testing set for the CNNs. This section provides the results about the two experiments to provide a clearer overview of the disparity between the different performance.

4.1. First Experiment

In this experiment the LSTM process input signals of fixed length with 200 hidden layers and *adam* optimizer with 0.001 of learning rate. For the CNNs, instead of the input size change depending on the input size of each CNN, the optimizer is *sgdm* with 0.001 of learning rate. The number of epochs can change based on network performance. Table1 table shows the PVC detection results ordered from best to worst based on the F1-score.

4.2. Second Experiment

Instead of the only difference in this experiment is that the LSTM has 100 hidden layers. The2 table shows the

Table 1: First experiment results.

Networks	F1	Acc	Se	Sp	PPV	AUC
RN-50	99.16	99.85	99.09	99.93	99.23	99.50
GN	98.62	99.76	97.83	99.95	99.43	98.88
AN	98.53	99.74	98.39	99.87	98.66	99.13
I-V3	98.29	99.70	98.74	99.79	97.85	99.26
GN(NP)	96.09	99.32	96.63	99.57	95.56	98.10
LSTM	91.28	98.53	89.18	99.41	93.48	94.29

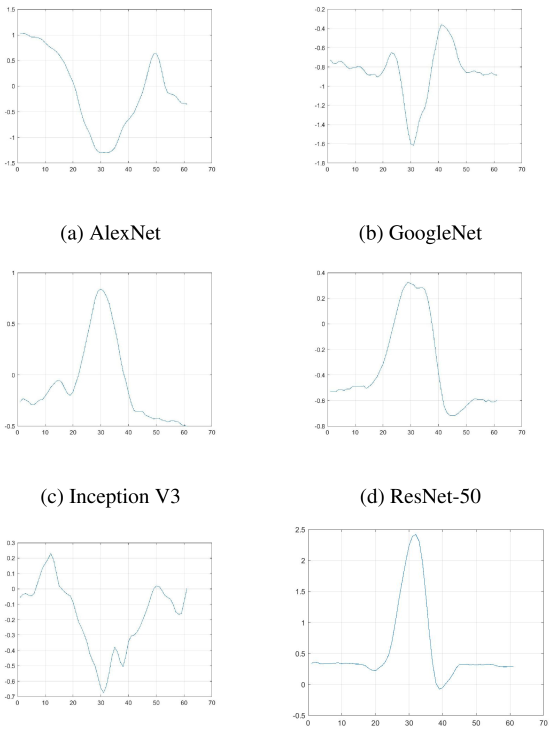
results of the PVC detection for this experiment, ordered, also in this case, from the best to the worst based on the F1 score.

Table 2: Second experiment results.

Networks	F1	Acc	Se	Sp	PPV	AUC
I-V3	98.77	98.77	98.53	99.02	99.01	98.77
GN	98.66	98.67	97.97	99.37	99.36	98.66
RN-50	98.42	98.42	98.39	98.46	98.46	98.42
AN	98.29	98.28	98.53	98.04	98.05	98.28
GN(NP)	98.06	98.07	97.55	98.60	98.58	98.07
LSTM	98.03	98.07	97.65	98.48	98.42	98.18

5. Discussion

The experiments in the previous section showed excellent results obtained with high accuracy both on the entire dataset and on the balanced dataset. The CNNs compared to the LSTM record higher performances, also in the case of NP GoogleNet, even if only slightly. The reason for this is that the CNNs used worked in different contexts and modes, but considering the little difference between the various performances, both CNNs and LSTM provides reliable PVC detection while maintaining stable performance. As noted in Section 4, the ResNet-50 was the best classifier in the first experiment achieving 99.8% of Acc and F1, Sp and AUC of 99.16%, 99.93% and 99.50%, while the Inception V3 was the best classifier in the second experiment with 98.8% of Acc, 98.8% of F1, 98.02% of Sp and 98.8% of AUC. Based on the accuracy, statistical significance tests were also used to compare the performance of the classifier, the Chi-square, which more precisely demonstrated the discrepancies between the results. The most interesting information can, however, be extrapolated from the study of the confusion matrix. The analysis of the *False Negative* and the *False Positive* images allows us to perform a ***"failure analysis"*** to explain where and why the classifiers made incorrect classifications and it should be noted that the same typologies of images classified incorrectly recur frequently in the two different experiments and also that these images are also often classified in the same wrong class.
More particularly regarding the ***False Positives***, there is often the presence in the various experiments of images with only long downward curves and/or small curves before and after the QRS. The Fig. 2 shows examples of these images that mistakenly lead the classifier to think that they were PVC records but they are not.



(a) AlexNet

(b) GoogleNet

(c) Inception V3

(d) ResNet-50

(e) NP GoogleNet

(f) LSTM

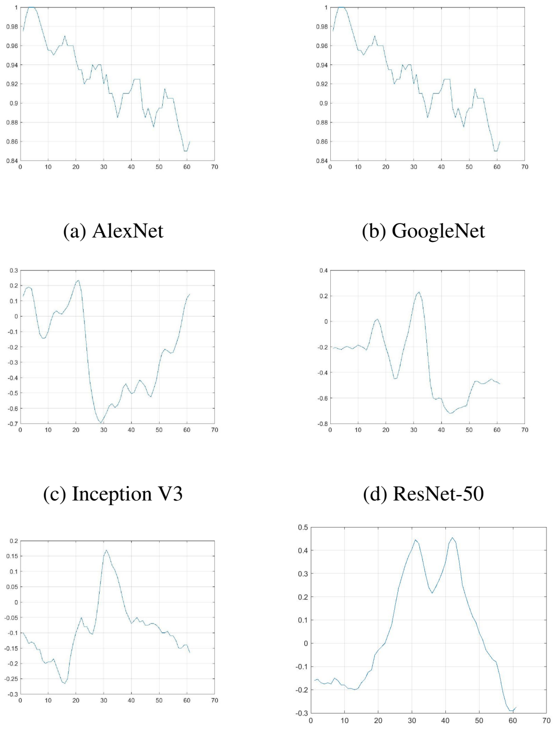
Figure 2: False Positives

On the other hand, as regards ***False Negatives***, in all experiments there are recurring images with a lot of noise (double curves or with many steps) and for these reasons the classifier, failing to identify them as non-PVC, classifies them as PVC. Fig. 3 shows examples of these images.

The major limitations of the project, however, were that the focus was only on PVC detection and not on identification of other heart rhythm problems but also the use of repeated measures from the same subject, yet the methods used provide reliable PVC detection diagnosis with high performance in both experiments.

6. Conclusion

Five deep neural networks were proposed and evaluated on the MIT-BIH Arrhythmia Database for the ECG rhythm evaluation, specifically for PVC detection techniques. In



(a) AlexNet

(b) GoogleNet

(c) Inception V3

(d) ResNet-50

(e) NP GoogleNet

(f) LSTM

Figure 3: False Negatives

recent years, deep learning has been shown to be more and to improve diagnostic quality to detect heart rhythms compared to other techniques, even with images of outpatient ECGs produced at half resolution [9], this research shows interesting results with images respecting raw data even when there was noise. Finally, the research questions outlined in the Section 1 can be answered.

1. What is the performance of the different deep learning models for PVC detection? For almost all models, the performance is higher than 97%, except for the LSTM. The performance of CNN models were greater than the performance of the LSTM model, according a Chi-square test ($p < 0.05$).

2. What type of deep learning model is the best classifier for detecting PVCs, such as CNNs or LSTM? The best classifier in the first experiment was ResNet-50, then the best classifier in the second experiment was Inception V3. The results suggest that training using the original time series sample data with an LSTM does not outperform a CNN trained using images. Perhaps it may hinder the performance.

3. What is the impact of pre-training on the performance of PVC detection deep learning models? The statistical analysis has shown that the pre-trained models perform better than non pre-trained models ($p < 0.05$).

4. What insights failure analysis provide into the false positive and false negative errors made by the classifier? The analysis of False Positive and False Negative to conduct a ”failure analysis” has shown interesting information to understand when and why the classifier made incorrect classification.

The detection of PVC is an environment where studies are still required to improve accuracy and to develop new techniques.

Acknowledgments

This work was supported by the Eastern Corridor Medical Engineering Centre that is funded by the European Union’s INTERREG VA Programme and managed by the Special EU Programmes Body (SEUPB)

References

[1] Talbi M, Charef A. Pvc discrimination using the qrs power spectrum and self-organizing maps. Computer Methods and Programs in Biomedicine Jun. 2009;94(3):223–231.

[2] Eugenio P. Frequent premature ventricular contractions — an electrical link to cardiomyopathy. Cardiol Rev Jul-Aug. 2015;23(4):168–72.

[3] Iwasa A, Hwa M, Hassankhani A, Liu T, Narayan SM. Abnormal heart rate turbulence predicts the initiation of ventricular arrhythmias. Pacing Clin Electrophysiol Nov. 2005; 28(11):1189–97.

[4] Sahoo S, Dash M, Behera S, Sabut S. Machine learning approach to detect cardiac arrhythmias in ecg signals: A surveys. IRBM Jan. 2020;.

[5] Liu Y, Huang Y, Wang J, Liu L, Luo J. Detecting premature ventricular contraction in children with deep learning. J Shanghai Jiaotong Univ Sci Mar. 2018;23:66–73.

[6] Zhou F, Jin L, Dong J. Premature ventricular contraction detection combining deep neural networks and rules inference. Artificial Intelligence in Medicine Jun. 2017;79:42–51.

[7] Jun TJ, Park HJ, Minh NH, Kim D, Kim Y. Premature ventricular contraction beat detection with deep neural networks. EEE International Conference on Machine Learning and Applications ICMLA Nov. 2016;859–864.

[8] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature May 2015;521(7553):436–44.

[9] Brisk R, Bond RR, Banks E, Piadlo A, Finlay D, McLaughlin J, David M. Deep learning to automatically interpret images of the electrocardiogram: Do we need the raw samples? Journal of Electrocardiology Oct. 2019;57:S65–S69.

Address for correspondence:

Fabiola De Marco
Via San Paolo, Gioi, 84056, SA, Italy
f.demarco10@studenti.unisa.it